Essays in Microeconometrics

Dissertation
for the Faculty of Economics, Business Administration and Information Technology of the University of Zurich

to achieve the title of
Doctor of Philosophy
in Economics

presented by

Kevin E. Staub
from Horgen

approved in April 2011 at the request of
Prof. Dr. Rainer Winkelmann
Prof. Dr. Josef Zweimüller

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorises the printing of this Doctoral Thesis, without thereby giving any opinion on the views contained therein.

Zurich, April 6, 2011

Chairman of the Doctoral Comittee: Prof. Dr. Dieter Pfaff

Acknowledgements

During the completion of this dissertation, I worked as a research assistant for Rainer Winkelmann at the Chair for Statistics and Empirical Economic Research in Zurich. I find it hard to imagine a more fertile environment for me to have written this thesis in; I benefitted enormously from innumerable stimulating conversations with Rainer Winkelmann, whom I cannot thank enough for his patience, encouragement and generosity in sharing knowledge.

I am also deeply indebted to Stefan Boes, whom I owe the answers to myriads of questions. I further would like to thank Steven Stillman for many extended late-afternoon conversations.

Peter Egger and Mario Larch had the kindness of inviting me to CES in Munich for two short research visits in 2008 and 2009; I thank them and members of their group for their welcoming hospitality. I had the privilege of working with Peter and Mario, and our collaboration has been enormously productive and inspiring to me.

I have participated in many lively discussions on econometrics (and other subjects) with Gregori Baetschmann, Timo Boppart and Raphael Studer. Arguing in this tough forum has always been an immense pleasure to me, and I look forward to many more sessions.

I thank Adrian Bruhin, Alejandra Cattaneo, Philippe Mahler and Christian Kascha for contributing to the excellent work atmosphere at the Chair, and Josef Zweimüller for co-advising this thesis.

Finally, I want to thank Susanita J. Méndez for reading and commenting on dozens of my drafts, and for her constant and loving support.

Contents

List of Tables

List of Figures

Chapter 1

Dissertation overview

The chapters in this dissertation deal with a diverse set of current topics in microeconometrics. The outcome variables in the econometric models studied are all limited dependent variables (LDV), i.e. their domain is only a subset of the real line. Beyond that, the LDV discussed here are quite heterogeneous, including both quantitative and qualitative, as well as discrete and continuous outcome variables. The topics' range is reasonably broad, too. It includes identification, hypothesis testing, and estimation. The issues are investigated analytically and by Monte Carlo simulations, and put to practice in applications. While the chapters are quite independent from each other, there are some common preoccupations which surface throughout this dissertation. Before expanding on these, I will begin by giving an overview of the contents of the chapters.

Chapter 2 investigates power and size of some tests for exogeneity of a binary explanatory variable in count models. Potentially endogenous binary regressors constitute a research strand of great relevance in empirical economics, because it is an often-encountered situation in the context of policy evaluation in non-experimental settings. The object of interest is the elasticity of the expected outcome with respect to the treatment, the binary policy. In applications, older contributions to the literature typically assumed exogeneity of treatment, while more recent work stresses possible non-random assignment to and self-selection into treatment which violate the exogeneity assumption. By using instrumental variable techniques these newer strands of the literature not seldom reach conclusions that differ drastically from older efforts. Tests for exogeneity constitute an important part of the argument of this newer literature because rejecting the exogeneity hypothesis makes it more credible to attribute the differences in results to the presence of endogeneity.

Chapter 2 compares such exogeneity tests in the context of count dependent variables by conducting extensive Monte Carlo simulations. The tests under consideration are Hausman contrast tests as well as univariate Wald tests, including a new test of notably easy implementation. This new test is based on the inclusion as an auxiliary regressor of the

generalized residual. Performance of the tests is explored under model misspecification and under different conditions regarding the instruments. The results indicate that often the simpler tests outperform tests that are more demanding to estimate; this is especially the case for the new test. The chapter also warns against the popular practice of using these tests as pretests to decide whether it is necessary to use instrumental variables. Simulation evidence strongly suggests that such a practice has devastating consequences on inference about the object of interest.

Chapter 3, written jointly with Rainer Winkelmann, considers estimation of LDV models where the outcome is a zero-inflated count. Applications of zero-inflated count data models have proliferated in empirical economic research. The reason is that count data used by social scientists often contain much more observations with an outcome of zero than standard count models would predict. Zero-inflated counts can accommodate arbitrary fractions of zeros and offer the convenient interpretation of resulting from the pooling of two unobservable subpopulations' outcomes, one having a standard count distribution and another one having a degenerated distribution with only outcome zero. In the literature, zero-inflated count models have always been estimated by Maximum Likelihood. This requires the full specification of the standard count part of the outcome variable. Prominent examples are the zero-inflated Poisson and the zero-inflated negative binomial model. However, the Maximum Likelihood estimators of these models are not robust to misspecification. Chapter 3 shows that, in contrast, simple Poisson Quasi-Likelihood estimators are consistent even in the presence of excess zeros and they do not require specifying the count distribution completely. The advantages of the Poisson Quasi-Likelihood approach are illustrated in a series of Monte Carlo simulations and in an application to the demand for health services.

Chapter 4 considers models for Tobit-type dependent variables. These are outcomes with domain over the nonnegative real numbers and positive probability mass at zero. When evaluating policies on such variables, a central object of interest is the decompo-

sition of the average treatment effect into what is called the extensive and the intensive margin. The extensive margin contribution is the portion of the treatment effect due to individuals changing from zero to positive values of their outcome in response to the policy; the intensive margin is the contribution to the average effect of individuals which in the absence of the policy have a positive outcome. However, this chapter shows that the usual decomposition of the average treatment effect in these models used in the literature is generally incompatible with a causal interpretation. I propose a decomposition based on the joint distribution of potential outcomes which is meaningful in a causal sense. The difference between decompositions can be substantial and even produce diametrically opposed results, as shown in a standard Tobit model example. In a generalized Tobit application exploring the effect of reducing firm entry regulation on bilateral trade flows between countries, estimates suggest that using the usual decomposition would overstate the contribution of the extensive margin by around 15%.

Finally, Chapter 5, written jointly with Gregori Baetschmann and Rainer Winkelmann, considers a model for ordered LDV in the context of panel data with correlated individual-specific unobserved heterogeneity, a so-called fixed effects model. If neglected, this kind of heterogeneity causes estimators to be biased. In economics, such time-invariant unobserved characteristics are a major source of concern because agents' (time-invariant) preferences are likely to be correlated with choice variables which appear as regressors and outcomes in empirical models. The fixed effects ordinary least squares (OLS) estimator for panel data uses a straightforward transformation of the variables which gets rid of the fixed effects and estimates parameters consistently, but such easy transformations are not available for nonlinear models in general. The chapter re-examines existing estimators for the panel data fixed effects ordered logit model, proposes a new one, and studies the sampling properties of these estimators in a series of Monte Carlo simulations. There are two main findings. First, we show that some of the estimators used in the literature are inconsistent, and provide reasons for the inconsistency. Second, the new estimator is never outperformed

by the others, seems to be substantially more immune to small sample bias than other consistent estimators, and is easy to implement. The empirical relevance is illustrated in an application to the effect of unemployment on life satisfaction.

The central feature underlying all chapters is that the problems studied are motivated directly from widespread empirical practices (as opposed to, say, from a particular application). While examples and applications are drawn mainly from fields of economics, this primarily reflects tastes and education of the author(s). The methods presented in this dissertation are also relevant to social scientists in general. Indeed, the practices discussed here can also be found in the empirical literatures of sociologists and of political scientists, for instance.

A second common point is that the proposed solutions are characterized by simplicity. While simplicity can be defended as a scientific value worth pursuing on its own (cf. for instance, Keuzenkamp and McAleer, 1995, for a formal treatment), simple strategies are also more likely to be of practical relevance, thus linking them back to the first feature. Simple solutions mean here principally that their implementation is straightforward and can be achieved either with existing software or by slightly modifying existing software. However, the solutions are also simple in a conceptual sense, which hopefully contributes to clarifying the issues at stake. For instance, Chapter 3 argues that since most researchers are interested in effects on the conditional expectation function (CEF) of zero-inflated count models, only this feature of the model should be exploited for estimation.

Since traditionally LDV models have been estimated by Maximum Likelihood (ML) and it still continues to be the most common approach, the chapters are mainly cast in the Maximum Likelihood framework, too. ML is by no means the only possible approach to estimating LDV models, and alternatives include OLS, generalized method of moments and nonparametric estimation. There are some contact points to some of these methods throughout the thesis. E.g., the pseudo-likelihood estimators presented in Chapter 3 are

method of moments estimators, and Chapter 4 includes a discussion on nonparametric identification; Chapter 2 compares the performance of tests based on method of moments estimators to tests based on ML estimators.

The point of which approach is best suited to estimate LDV models merits a little bit more expansion, as this question has motivated extensive discussions in the past, and continues provoking debates at present; positions vary widely, and it is unlikely that a definite consensus will ever be achieved. Economists' empirical workhorse is the OLS estimator, which is a consistent estimator of a linear CEF. With regards to LDV models, it has been argued that OLS estimation is inappropriate because (a) the CEF of these models is limited, and OLS estimation might yield out-of-range predictions; (b) the CEF of these models is nonlinear in general; (c) the CEF might not be defined for these models (e.g., for ordered LDV as in Chapter 5); and (d) conditional probabilities, not the CEF, might be the object of interest (cf. Cameron and Trivedi, 2005, Winkelmann and Boes, 2008). Modern responses to objections (b)—(d) to OLS estimation of LDV models emphasize the limited knowledge about data generating processes, in view of which the best (i.e. minimum mean square error) linear approximation property of OLS is very attractive: Provided a CEF exists, OLS will approximate it linearly; if the CEF does not exist, it is possible to define conditional probabilities, for each of which OLS estimation can be performed since every conditional probability coincides with a CEF. Given the prominent focus on CEF-effects in economics, modern advocates of OLS estimation question the importance of objection (a) (Angrist and Pischke, 2009).

These arguments seem plausible to me. However, a similar stance can be taken up on ML estimation, and view it as giving an approximation in case of misspecification. There is seldom a firm reason to preferring linear to other types of approximations, especially if it is known that the CEF cannot be linear. The body of literature on ML estimation of misspecified models dates back to White's (1982) seminal work; in a similar way as OLS minimizes the mean square error to the CEF, ML estimation minimizes the Kullback-Leibler

distance to the conditional density function. Chapter 2, for instance, finds acceptable performance of tests based on ML estimation of misspecified models. Furthermore, a broad class of ML estimators is known to retain consistency even if the model is misspecified (Gourieroux, Monfort and Trognon, 1984), and Chapter 3 studies such a case in detail.

A completely different argument in favor of the ML approach is based on conceptual clarification: The full parametric specification of a model can be helpful to understand possible mechanisms and causal pathways, irrespective of whether the chosen functional forms correspond to the data generating process or not. In Chapter 5, for instance, the Tobit model is used to illustrate the differences between a decomposition of causal effects used in the literature and a newly proposed decomposition. The fact that the parametric specification yields simple formulas that depend on few parameters with clear interpretations, helps fixing ideas and understanding the channels through which the differences arise. An exposition in a more general framework would risk drowning the central message in unnecessary details.

Finally, apart from practical relevance, simplicity and being rooted in the ML approach —features shared by all chapters— there are other themes which surface in some, although not in all chapters. Among these, three important topics represented each in two chapters are causal inference (Chapters 2 and 4), endogeneity (Chapters 2 and 5) and model misspecification (Chapters 2 and 3).

References

Angrist, Joshua D. and Jörn-Steffen Pischke (2009), *Mostly Harmless Econometrics*, Princeton University Press.

Cameron, A. Colin and Pravin K. Trivedi (2005), *Microeconometrics*, Cambridge University Press.

Gourieroux, Christian, Alain Monfort and Alain Trognon (1984), "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, **52**, 681-700.

Keuzenkamp, Hugo A. and Michael McAleer (1995), "Simplicity, scientific inference and econometric modelling", *Economic Journal*, **105**, 1-21.

White, Halbert (1982), "Maximum likelihood estimation of misspecified models", *Econometrica*, **50**, 1-25.

Winkelmann, Rainer and Stefan Boes, 2009, *Analysis of Microdata*, second edition, Springer.

Chapter 2

Simple tests for exogeneity of a binary explanatory variable in count data regression models

This chapter has been published in *Communications in Statistics – Simulation and Computation*, **38**(9), pp. 1834-1855.

Acknowledgements: The author wishes to thank João M.C. Santos Silva and an anonymous referee for helpful comments on this article. Special thanks to Rainer Winkelmann for extensive discussion and advise which significantly improved this article. Any remaining errors are the author's sole responsability.

2.1 Introduction

This article is concerned with inference about endogeneity caused by a binary variable in count data models. Unlike the case with a continuous endogenous regressor, such models cannot be consistently estimated by two-stage residual-inclusion procedures, making it necessary to use other estimation techniques. For instance, nonlinear instrumental variables estimation as introduced by Mullahy (1997) is general enough to be applicable irrespective of the binary nature of the endogenous regressor, and can therefore be used to conduct Hausman tests of endogeneity. If the focus is solely on testing exogeneity, however, easily implementable two-stage residual-inclusion also provides a valid test which was first proposed by Wooldridge (1997). Furthermore, if the researcher is willing to introduce parametric assumptions about the error structure of the model (Terza, 1998), significant efficiency gains might be exploited and alternative tests for exogeneity can be implemented.

Despite its rather specific nature, estimation of count data models with a potentially endogenous dummy variable is very common in the empirical economics literature, and with estimation routines for this models becoming available in statistical software packages[1] the number of applications is bound to increase further. Earlier examples of count data models with an endogenous dummy variable include Windmeijer and Santos Silva (1997), who study the effect of a binary measure of self-reported health on the number of physician consultations; Terza (1998) who investigates the impact of vehicle ownership on the number of recreational trips; and Kenkel and Terza (2001) who analyze how physician advice affects the consumption of alcoholic drinks. To cite just a few, more recent work studies whether educational attainment decreased women's fertility (Miranda, 2004), or if U.S. residence of mexican women influenced their relationship power as measured by the number of less egalitarian responses to a questionnaire (Parrado, Flippen and McQuiston, 2005). The model has also been used to test for possible endogeneity of the mechanism

[1] E.g., there are routines for both Mullahy's (1997) NLIV/GMM estimator and Terza's (1998) full information maximum likelihood estimator in STATA. See Nichols (2007) and Miranda (2004), respectively.

to price initial public offerings (bookbuilding or auction) in a regression on the number of buy recommendations for a company (Degeorge, Derrien and Womack, 2007). Quintana Garcia and Benavides Velasco (2008) investigated if an increase of diversification in firm technology lead to a higher number of patents.

The model has also been the subject of more theoretically-oriented work, which developed semiparametric procedures to estimate the model under less stringent assumptions (e.g. Romeu and Vera-Hernandez, 2005; Masuhara, 2008); a Bayesian version of the model is analyzed in Kozumi (2002). However, since the impact of these developments on applied work is more modest, and given that the focus of this article is on tests for exogeneity that are relevant for applied empirical practice, the analysis will be limited to exogeneity tests obtained under more widespread –if more restrictive– model assumptions.

Below, various tests for exogeneity in a count data model with a binary endogenous regressor are presented and their performance is compared in small and moderately-sized samples through Monte Carlo simulation. This article is restricted to the just-identified case with one instrument. As a benchmark, the Hausman test that contrasts efficient and consistent estimates is evaluated against various univariate Wald tests based on an estimated parameter that captures the degree of endogeneity. Among them, a new test of particularly easy implementation is presented. The tests are assessed with regards to sensitivity to instrument strength and to mild and moderate model misspecification of the data generating process. A key result of interest to practitioners is that, overall, the two most easy-to-implement tests, including the new test, displayed very acceptable empirical size and power properties among the presented tests, often outperforming the other tests.

Frequently endogeneity tests are conceived as pretests to decide whether a model estimated with an estimator that is consistent under endogeneity can be re-estimated with a more efficient estimator that is only consistent under exogeneity. However, recent work by Guggenberger (2008) in a linear IV model context demonstrates that using a Hausman pretest can be devastating for inference on second stage tests. Thus, further simulations

are performed to address the question of how exogeneity pretests affect inference about the effect of the potentially endogenous binary variable in count data models. Here, the results turn out to be less encouraging, as severe size distortions suggest that researchers should refrain from using these exogeneity tests as pretests.

The rest of the article is organized as follows. Section 2.2 presents the model under consideration. The tests for exogeneity are introduced in the next section. The design of the Monte Carlo experiment and its results are discussed in section 2.4, while section 2.5 contains some conclusions.

2.2 Count data regression models with a potentially endogenous binary variable

The model considered here will be a model for a count dependent variable, y, whose mean, conditional on a vector of observed explanatory variables x, a binary variable d and an unobserved error component ε, is an exponential function of a linear index of (x, d, ε):

$$\mathrm{E}(y|x, d, \varepsilon) = \exp(x'\beta + \beta_d d + \varepsilon) \tag{2.1}$$

Concentrating the analysis to this class of models means that the conclusions of this article are relevant to a wide range of applied work, since both Poisson and Negative Binomial regression, the two most extensively used count model estimators, fall by default into the class defined in (2.1)[2]. Note that including the error term ε in the exponential function as opposed to additively outside the function corresponds to the interpretation of ε as further variables that affect the expectation of y (but that are unobservable to the econometrician) and should be treated symmetrically to the observed variables[3].

[2]Evidently, exponential conditional mean functions are not limited to count data, and many of the procedures and results discussed here are in principle applicable to continuous data as well.

[3]An alternative justification for this representation is by means of the interpretability of the model in terms of ceteris paribus marginal effects (cf. Winkelmann, 2008, p. 160).

If the regressors x and the dummy variable d are statistically independent from ε, the conditional expectation function (2.1) marginal of ε is

$$\mathrm{E}(y|x,d) = \exp(x'\beta + \beta_d d)\,\mathrm{E}[\exp(\varepsilon|x,d)] = \exp(x'\beta^* + \beta_d d), \tag{2.2}$$

assuming that the mean of $\exp(\varepsilon)$ is constant and that x includes a constant first element, as then β^* is equal to β but with first element shifted by $\ln \mathrm{E}[\exp(\varepsilon)]$ (cf. Windmeijer and Santos Silva, 1997). Note that assuming zero correlation between regressors and errors as in the linear case is not sufficient for (2.2) to hold, as this does not warrant that $\mathrm{E}[\exp(\varepsilon)|x,d] = \mathrm{E}[\exp(\varepsilon)]$.

Equation (2) represents the case of exogeneity, and efficient estimation of the model depends on the distribution of ε and of $y|x,d,\varepsilon$. For instance, with the latter being Poisson-distributed, if ε is distributed as log-normal or exp-gamma, then the resulting models marginal of ε are the Poisson-log-normal and the negative binomial regression model, respectively. However, because of its robustness to distributional misspecification and easy implementation, it is very common to give up full efficiency and estimate models satisfying (2.2) by Poisson pseudo maximum likelihood (cf. Wooldridge, 1997), which yields consistent estimates of (β^*, β_d) irrespective of the distribution of ε. Nonlinear least squares estimation is also consistent, but is less frequently encountered in the count data context as it neglects the count nature of the dependent variable. Consistency up to the first element does not hold in general for nonlinear models but is a specific consequence of the multiplicative separability of linear combinations in the exponential function.

For continuous elements of x, the parameters β have the interpretation of (semi-)elasticities with respect to the conditional expectation function (CEF), i.e. for the kth regressor

$$\frac{\partial \mathrm{E}(y|x,d)/\mathrm{E}(y|x,d)}{\partial x_k} = \beta_k$$

while for discrete regressors, as for instance the binary variable of interest here, direct interpretation of the coefficients is only suitable as an approximation to the discrete partial

effect $\exp(\beta_d) - 1$. Note that for both marginal and discrete partial effects as well as for predictions of CEF, inconsistent estimation of the first element of β is inconsequential[4].

The binary variable d is endogenous in model (2.1) whenever it is not statistically independent from ε and, thus, the second equality in (2.2) does not hold. Estimation of the model neglecting endogeneity yields inconsistent estimates of all parameters, even when the regressors are orthogonal. To pin down the source of this dependence one can recur to modelling d as

$$d = \begin{cases} 1 & \text{if } z'\gamma \geq v \\ 0 & \text{if } z'\gamma < v \end{cases} \tag{2.3}$$

where z is a vector of observable variables, possibly including at least some elements from x, and the unobserved error component v follows some joint distribution with ε from (2.1). Terza (1998) proposed to specify the distribution of $(\varepsilon, v)'$ conditional on the exogenous variables (x, z) as bivariate normal according to

$$\begin{pmatrix} \varepsilon \\ v \end{pmatrix} \Big| \, x, z \sim Normal \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right] \tag{2.4}$$

which defines a probit model for (2.3). Also, statistical dependence is captured entirely by the correlation parameter $\rho \in [-1, 1]$ which yields independence whenever $\rho = 0$. Thus, the hypothesis of exogeneity can be stated as $H_0 : \rho = 0$ with alternative $H_1 : \rho \neq 0$ corresponding to endogeneity.

2.3 Tests for exogeneity

The most widely used test for exogeneity is probably the Hausman test, since it is applicable in a vast number of situations. In the context of the model discussed here, it has the advantage that it does not require assumption (2.4). After shortly discussing Hausman

[4]While the partial effects do not depend on the first element of β, predictions of CEF are consistent because $x'\hat{\beta}^*$ is consistent for $x'\beta + \ln \mathrm{E}[\exp(\varepsilon)]$.

tests, the exposition will turn to univariate Wald tests, first presenting two tests based on Terza's (1998) full information maximum likelihood estimator and a more general two-stage method of moments estimator. Finally, two tests of particularly easy implementation are discussed, which also rely on estimation in two stages: a new test based on a first order approximation to the method of moments estimator and a residual inclusion estimator.

2.3.1 Hausman contrast tests

The Hausman test (Hausman, 1978) in its most general form contrasts two estimates obtained from different estimators. In the case of endogeneity, one of the estimators is consistent under both the null hypothesis (exogeneity) and the alternative (endogeneity) while the second estimator is inconsistent under the alternative but efficient (relative to any linear combination of the two estimators) under the null hypothesis. Then, denoting by $\hat{\beta}_C$ the consistent estimate and by $\hat{\beta}_E$ the efficient one, the Hausman test statistic is

$$h = (\hat{\beta}_E - \hat{\beta}_C)'[\mathrm{Var}(\hat{\beta}_C) - \mathrm{Var}(\hat{\beta}_E)]^{-1}(\hat{\beta}_E - \hat{\beta}_C) \quad \sim \quad \chi^2_j$$

with the degrees of freedom of the χ^2 distribution, j, being equal to the dimension of the β-vectors involved in h.

An early application of a Hausman test to count data models with endogeneity is provided by Grogger (1990), who suggested calculating the corresponding test statistic with estimates from Poisson ML and a nonlinear instrumental variables (NLIV) estimator based on an additive error to the CEF. However, this estimator is inconsistent under a multiplicative error defined implicitly as $\exp(\varepsilon)$ in (2.1) (Dagenais, 1999; Terza, 2006), and Mullahy's (1997) GMM estimator is therefore more appropriate to estimate $\hat{\beta}_C$. In the just-identified case studied here, this estimator is the NLIV based on the residual function $r \equiv y\exp(-x'\beta - \beta_d d) - 1$ which, given an appropriate instrument z, implies the moment condition

$$\mathrm{E}(r|z) = \mathrm{E}[\exp(\varepsilon) - 1|z] = 0$$

Thus, writing the NLIV estimate of β_d as $\hat{\beta}_d^{NLIV}$ and the corresponding Poisson PML estimate as $\hat{\beta}_d^{PPML}$, a Hausman test for exogeneity can be based on the test statistic

$$h^1 = \frac{(\hat{\beta}_d^{PPML} - \hat{\beta}_d^{NLIV})^2}{\text{Var}(\hat{\beta}_d^{NLIV}) - \text{Var}(\hat{\beta}_d^{PPML})} \quad \sim \quad \chi_1^2 \tag{2.5}$$

Sometimes this Hausman test is implemented by additionally including all elements of β in the contrast, but both Creel (2004) and Chmelarova (2007) find that h^1 outperforms the full-β-version of the test in finite samples.

The denominator of h^1 results as a special case of the variance of a difference of estimates when the minuend is the efficient estimator, as then $\text{Cov}(\beta_E, \beta_C) = Var(\beta_E)$ (Hausman, 1978). There are two routes of potentially improving on h^1. The first would be to specify the distribution of ε and then calculating the corresponding ML estimator. For instance, if (2.4) holds, the model for y conditional on observables is a Poisson-log-normal (PLN) mixture. As the PLN estimator is efficient relative to the Poisson estimator in this model, a Hausman test statistic calculated by substituting the PPML estimates by PLN equivalents could perform better:

$$h^2 = \frac{(\hat{\beta}_d^{PLN} - \hat{\beta}_d^{NLIV})^2}{\text{Var}(\hat{\beta}_d^{NLIV}) - \text{Var}(\hat{\beta}_d^{PLN})} \quad \sim \quad \chi_1^2$$

A second procedure in the vein of Weesie (1999) and Creel (2004) is to estimate $\text{Cov}(\beta_E, \beta_C)$ directly instead of relying on the simplification under asymptotic efficiency[5]. This implies to rewrite the two optimization problems of the Poisson PML and the NLIV as a joint problem by stacking PPML's first order conditions and the moment conditions of NLIV. The resulting test statistic is

$$h^3 = \frac{(\hat{\beta}_d^{PPML} - \hat{\beta}_d^{NLIV})^2}{\text{Var}(\hat{\beta}_d^{PPML}) + \text{Var}(\hat{\beta}_d^{NLIV}) - 2\text{Cov}(\hat{\beta}_d^{PPML}, \hat{\beta}_d^{NLIV})} \quad \sim \quad \chi_1^2$$

If the errors follow a bivariate normal distribution, all three tests are asymptotically equivalent. If not, h^2 is inconsistent, but h^1 and h^3 retain their consistency. The perfor-

[5]Creel's (2004) approach is optimal GMM, while Weesie (1999) does not use a second step weighting matrix. Clearly, in the just identified case under consideration both amount to the same as the choice of the weighting matrix does not affect the estimates.

mance of the two additional variants relative to h^1 is less clear in finite samples. For h^3 the potential gains depend crucially on the small sample properties of the covariance estimator. Likewise, for h^2 to outperform h^1 the higher precision of PLN relative to Poisson – which is an asymptotic result – needs to be visible enough in finite samples.

2.3.2 Wald tests

There are alternatives to the Hausman contrast test for exogeneity. For instance, in the linear IV model, estimating a reduced form for the endogenous variable in order to obtain residuals which can be plugged into the structural equation leads to an asymptotically equivalent test for endogeneity (Hausman, 1978). Monte Carlo simulations in Chmelarova (2007) show that Wald versions of the Hausman test often have better properties than the contrast version under a series of different conditions. However, the endogeneity in count data models in Chmelarova (2007) concerns continuous regressors, so that the residual inclusion technique is consistent. Residual inclusion in the framework discussed presently with an endogenous dummy, on the other hand, yields inconsistent estimates[6]. Nevertheless, a number of consistent Wald tests are available.

First, Wooldridge (1997) suggests that while the procedure yields inconsistent estimates, the test based on residual inclusion is consistent. Second, if one is willing to impose (2.4) and a distributional assumption for $y|x, d, \varepsilon$, one can recur to Terza's (1998) maximum likelihood estimator, which explicitly estimates the correlation coefficient of the bivariate normal distribution so that the hypothesis $\rho = 0$ can be tested directly. Relaxing the distributional assumption on the dependent variable still allows to estimate a scaled version of ρ based on (2.4), which can be used to test for endogeneity. Last, following the literature

[6]Terza, Basu and Rathouz (2008) show that residual inclusion in nonlinear models is inconsistent in general. Discussions of consistency of residual inclusion in Poisson PML models with continuous endogenous regressors and inconsistency with binary regressors can be found inter alia in Wooldridge (1997) and Winkelmann (2008).

on inference using local approximations (cf. Chesher, 1991; Gourieroux and Visser, 1997), one can derive a test based on the inclusion of a generalized residual in the structural equation. While the second strategy yields consistent estimates for β_d under the alternative, the first and last do not. Their advantage, however, lies in their easy implementation, since only a standard Poisson regression is needed to carry out these tests.

Full information maximum likelihood and two-stage method of moments estimation

Assuming that (2.4) holds and that $y|x, d, \varepsilon$ follows a Poisson distribution with expectation (2.1), maximum likelihood estimation of the joint model proceeds by maximizing the sample log-likelihood function $\mathcal{L}(\beta_d, \beta, \gamma, \rho, \sigma) = \sum_{i=1}^{n} \log f(y_i, d_i|x_i, z_i)$, with $f(\cdot)$ denoting the probability density function, which given the assumptions is equal to (Terza, 1998)

$$\begin{aligned} f(y, d|x, z) &= \int_{-\infty}^{\infty} f(y|d, x, z, \varepsilon) \times f(d|x, z, \varepsilon) \times f(\varepsilon|x, z)\, d\varepsilon \\ &= \int_{-\infty}^{\infty} \exp(\lambda)\lambda^y (y!)^{-1} \times \Phi^*(\varepsilon)^d (1 - \Phi^*(\varepsilon))^{1-d} \times \sigma^{-1}\phi(\varepsilon/\sigma|x, z)\, d\varepsilon, \end{aligned}$$

where $\lambda \equiv \exp(x'\beta + \beta_d d + \varepsilon)$ and $\Phi^*(\varepsilon) \equiv \Phi\left(\frac{z'\gamma + \frac{\rho}{\sigma}\varepsilon}{\sqrt{1-\rho^2}}\right)$; $\Phi(\cdot)$ and $\phi(\cdot)$ denoting the cdf and pdf of the standard normal distribution, as usual. While the expression for $f(y, d|x, z)$ has no closed form solution, it is possible to approximate it through Gauss-Hermite quadrature. Given the ML estimate $\hat{\rho}$, the null hypothesis $H_0 : \rho = 0$ is tested constructing the t-statistic

$$t^1 = \frac{\hat{\rho} - 0}{s.e.(\hat{\rho})} \sim N(0, 1) \tag{2.6}$$

with $s.e.(\hat{\rho})$ indicating any usual asymptotically valid ML standard error of $\hat{\rho}$.

Terza (1998) also suggested a two stage estimation of this model which leaves $f(y|d, x, z, \varepsilon)$ unspecified. While the relaxation of assumptions is rather moderate as bivariate normality of the errors is maintained, the gains of such a procedure lie mostly in increased computa-

tional stability[7]. Consider (2.1) under assumption (2.4):

$$\begin{aligned}
\mathrm{E}(y|x,d) &= \exp(x'\beta + \beta_d d)\,\mathrm{E}(\exp(\varepsilon)|x,d) \\
&= \exp(x'\beta + \beta_d d)\,\exp\left(\frac{\sigma_\varepsilon^2}{2}\right)\left[d\frac{\Phi(\theta + z'\gamma)}{\Phi(z'\gamma)} + (1-d)\frac{1-\Phi(\theta+z'\gamma)}{1-\Phi(z'\gamma)}\right] \\
&\equiv \exp(x'\beta^* + \beta_d d)\,\psi(\theta,\gamma;z)
\end{aligned}$$

with $\theta = \sigma\rho$. To estimate this model in stages, first a probit regression is performed to obtain estimates of γ, so that in a second stage estimation optimization proceeds with respect to (β, β_d, θ). Terza's (1998) suggestion is to implement the second stage as nonlinear least squares (NLS), or as nonlinear weighted least squares (NWLS) if the researcher wishes to incorporate a priory knowledge of the distribution of $y|d,x,z,\varepsilon$.

In the present work, however, the second stage estimation will also be implemented as a Poisson pseudo-ML regression, i.e., estimates of (β, β_d, θ) are obtained by maximizing a pseudo-log-likelihood function of the Poisson distribution with expectation $\tilde{\lambda} \equiv \exp(x'\beta^* + \beta_d d)\psi(\theta, \hat{\gamma}; z)$. This estimation strategy represents a compromise between NLS and NWLS, in the sense that it is bound to be more efficient for count data than NLS since it takes account of the inherent heteroskedasticity characteristic of count data[8], while it avoids the computational difficulty of the more efficient NWLS procedure.

With an estimate of θ, the pertinent t-statistic of the test with null hypothesis $H_0 : \theta = 0$ is

$$t^2 = \frac{\hat{\theta} - 0}{s.e.(\hat{\theta})} \sim N(0,1) \tag{2.7}$$

[7]An important aspect of leaving $f(y|d,x,z,\varepsilon)$ unspecified is that it broadens the class of models this estimator is applicable to to other non-counts exponential CEF models. See, for instance, Egger et al. (2009) who apply such a model to bilateral trade.

[8]The argument for Poisson pseudo-MLE against NLS is presented extensively by Santos Silva and Tenreyro (2006) in the context of non-count exponential CEF models.

Generalized residual inclusion

It is possible to approximate the estimation of the two-stage method described above without the need of estimating a Poisson regression with mean $\tilde{\lambda}$, which in general requires some extra programming as standard econometric software usually only allow to specify variables entering a linear index in the exponential function. This is related to Greene's (1995, 1998) work in the context of sample selection in count data models. The starting point of this approximation is again (2.1) under assumption (2.4), which written separately for the two possible outcomes of d is

$$\mathrm{E}(y|x,d=1) = \exp(x\beta^* + \beta_d d)\frac{\Phi(\theta + z'\gamma)}{\Phi(z'\gamma)} = \exp(x\beta^* + \beta_d)Q_1 \quad \text{and}$$
$$\mathrm{E}(y|x,d=0) = \exp(x\beta^*)\frac{1-\Phi(\theta + z'\gamma)}{1-\Phi(z'\gamma)} = \exp(x\beta^*)Q_0,$$

Taking logarithms of the endogeneity bias correction terms Q_0 and Q_1 allows to write them as part of the linear index in the exponential function. Furthermore, the first order Taylor series expansion of $\log Q_0$ and $\log Q_1$ around $\theta = 0$ is

$$\log Q_1 \approx \theta\frac{\phi(z'\gamma)}{\Phi(z'\gamma)} \quad \text{and} \quad \log Q_0 \approx \theta\frac{-\phi(z'\gamma)}{1-\Phi(z'\gamma)},$$

so that the second stage of the former estimator can be approximated by estimating a Poisson pseudo-ML regression with expectation

$$\mathrm{E}(y|x,d) \approx \exp(x'\beta^* + \beta_d d + \theta m), \quad \text{with} \quad m = d\frac{\phi(z'\gamma)}{\Phi(z'\gamma)} + (1-d)\frac{-\phi(z'\gamma)}{1-\Phi(z'\gamma)}$$

and replacing m with a consistent estimate $\hat{m}$ obtained with probit estimates $\hat{\gamma}$[9].

Estimates of m represent generalized residuals in the sense that the first order conditions in the estimation of γ in the reduced form are a set of orthogonality conditions between m and z. Orme (2001), who introduced the same local approximation in the context of a dynamic probit model, proposed testing for the presence of an endogenous initial condition by using the estimated coefficient on the generalized residuals, $\hat{\theta}$. The same procedure can

[9]This technique has also been used by Angrist (2001) to approximate a Tobit MLE.

be applied here, suggesting a new test for exogeneity in the present count data context: If $\rho = 0$ the approximation is exact, so that the pseudo-ML estimates of θ will be consistent under the null hypothesis of exogeneity and the test statistic t^2 in (2.7) can be used.

Residual inclusion

While a glance at the pertinent literature shows that many resarchers are comfortable with assumption (2.4), the test proposed in Wooldridge (1997) is consistent under weaker distributional assumptions as it does not require bivariate normality. It does, however, in contrast to the Wald tests considered so far, require instruments.

The residual inclusion estimation procedure consists in including residuals from the reduced form equation for the endogenous variable in the linear index of the second stage exponential CEF. The two key assumptions for consistency of this technique are independence of the reduced form residuals from the instruments and linearity of the CEF of ε given v. The linear CEF condition holds if, as considered so far, the error terms are bivariate normally distributed. However, independence of the residuals from the instruments is unlikely to hold in the binary case. Nevertheless, as pointed out by Wooldridge (1997), the procedure is still valid to test for exogeneity, since under the null hypothesis of d being exogenous the two assumptions on the errors need not hold as then the CEF reduces to (2.2). I.e., while the procedure does not yield consistent estimates, it does provide a valid Hausman-type Wald test for endogeneity.

Starting with assumption (2.4), the CEF of ε given v is $\mathrm{E}(\varepsilon|v) = \theta v$, with $\theta = \sigma\rho$ as before. Therefore, it is always possible to write $\varepsilon = \theta v + error$, with this error being independent of v by construction. Thus, the suggested test would proceed by replacing ε in (2.1) with $\theta v + error$ and conditioning y on x, d and v (instead of ε). That is, estimating

$$\mathrm{E}(y|x, d, v) = \exp(x'\beta + \beta_d d + \theta v)$$

by Poisson pseudo-ML, using $\hat{v} = d - \Phi(z'\hat{\gamma})$ for the unobserved v, where estimates for γ could be obtained from a probit regression or, alternatively, from other models for binary

dependent variables such as the linear probability model, which would produce residuals $\hat{v} = d - z'\hat{\gamma}$. Again, the null hypothesis of exogeneity is expressed as $\theta = 0$ and the test statistic t^2 can be used.

2.4 A Monte Carlo simulation study

To assess finite sample properties of the tests discussed in the previous sections, a Monte Carlo simulation experiment is conducted. Bearing in mind the known limitations of such an approach, special care has been placed on addressing a variety of issues concerning the performance of the tests under different conditions, such as moderate misspecification and unavailability of instruments, as well as suitability of the tests for pretesting. All programming has been written in GAUSS, pseudo-random number generators and other subroutines used were taken from GAUSS' libraries; code and a supplementary appendix containing more extensive results are available from the author on request.

2.4.1 Experimental design

Every reported simulation proceeded by drawing a random sample of size n from two independent standard normally distributed variables, x and z. Next, the errors ε and v were drawn from some joint distribution having 0 expectations and variance of v equal to 1. The endogenous binary variable, d was formed according to

$$d = \mathbb{1}(\gamma_z z + \gamma_x x + v \geq 0)$$

with $\mathbb{1}(\cdot)$ denoting the indicator function. Then, the conditional expectation of the count dependent variable y was constructed as

$$\lambda = \exp(-1 + 0.5x + d + \varepsilon)$$

so that, finally, y was obtained by random sampling from some count data distribution with expectation λ. Here the effect of the dummy on the expectation of y is $\exp(1) - 1 \approx 1.71$

which might seem above what can be expected in some empirical applications, but adherence to the unit coefficient on d can be defended on the grounds of comparability to other studies[10]. Sample sizes (n) considered were 200, 500 and 1'000. Results for larger samples are not reported as then differences between tests even out quickly and they converge to their asymptotic limits. Smaller samples, on the other hand, were not investigated as microeconometric applications of this model with less observations are unlikely to be encountered in practice. Most Monte Carlo simulations were replicated 10'000 times, the significantly more computing-intensive routines for the tests based on full information maximum likelihood (FIML) estimates were performed with 2'000 and 1'000 replications. All tests were performed at a nominal significance level of 5%. Different data generating processes were obtained by varying the values of the vector γ, the joint distribution of the errors and the distribution of $y|x, d, \varepsilon$.

By assigning different values to γ, the strength of the instrument was manipulated. While in the linear IV model the concentration parameter provides an unequivocal summary measure of instrument strength (cf. Stock, Wright and Yogo, 2002), there is no generic equivalent for nonlinear models. Trivially, the impact of the instrument is affected by the proportion of the variance of $(\gamma_z z + \gamma_x x + v)$ explained by $\gamma_z z$. Note that a given ratio can be obtained by either changing the variance of the error v with respect to the given variance of $(\gamma_z z + \gamma_x x)$, or by altering the relation $\text{Var}(\gamma_z z)/\text{Var}(\gamma_x x)$ with given relation of $\text{Var}(\gamma_z z + \gamma_x x)$ to $\text{Var}(v)$. While the two interventions amount to the same in the linear model, here results might differ.

The pdf $f(y|x, d, \varepsilon)$ was set to be either Poisson with mean λ or Negative Binomial I with mean λ and variance 2λ. With the exception of the test based on full information maximum likelihood, all tests should be invariant to the overdispersion introduced by the Negative Binomial I variant. The baseline specification for the error distribution was the

[10]Monte Carlo studies of count data models with unit coefficient on endogenous variables include Creel (2004), Romeu and Vera-Hernandez (2005) and Chmelarova (2007).

bivariate normal distribution given in (2.4) with values of ρ ranging from 0 to 0.95 for most experiments. To assess sensitivity to misspecification of (2.4), (ε, v) were also generated from a bivariate Gaussian copula with an exponential Gamma marginal distribution for ε and a standard logistic marginal for v, inducing a Negative Binomial model for y conditional on observables and a logit model for d. Finally, the tests were conducted with the errors following the same exp-Gamma and logistic marginals but with joint distribution determined through the Frank copula.

A table containing the descriptions of the precise data generating processes that were used in producing the results discussed below can be found in the appendix (cf. Table 2.6).

The next subsection discusses empirical size and power of the proposed tests under ideal assumptions on the data generating process, i.e. with assumption (2.4) holding. Next, the discussion centers on the tests that theoretically are able to identify exogeneity in the absence of instruments, assessing the goodness of their performance under this condition in the simulations. Results under misspecification of the data generating process are considered next, and the section closes considering the effect on the empirical size of tests on $\hat{\beta}_d$ after using endogeneity tests as pretests to choose between estimators for the model.

2.4.2 Empirical size and power

The first three columns of Table 2.1 contain simulation results for the empirical size of different tests for exogeneity with nominal size 5%. The table shows results for three different sample sizes of 200, 500 and 1'000 observations. The coefficients of the reduced form equation, γ_x and γ_z were set to $\sqrt{0.5}$ each, so that the ratio $\text{Var}((\gamma_z z + \gamma_x x)/\text{Var}(v)$ equalled 1. With 10'000 replications, a 95% confidence interval for the estimated size of tests is $[0.05 \pm 1.96\sqrt{0.05 \times 0.95/10'000}] \approx [0.046, 0.054]$.[11]

The first three rows contain the rejection frequencies of the exogeneity hypothesis for the Hausman tests with test statistics h^1, h^2 and h^3 discussed previously. The test that

[11]The corresponding confidence interval for 2'000 replications is approximately [0.405,0.595].

contrasts PPML estimates with the NLIV estimates (H1) performs better than the two other Hausman tests. While underrejecting the true null hypothesis with 200 observations, H1 displays correct size for larger samples, while H2, which uses PLN estimates instead of PPML, underrejects slightly even for the largest sample. The test H3, which attempts to improve on H1 by estimating the covariance from the data instead of relying on the asymptotic simplification, has a serious underrejection problem for all sample sizes considered. Since estimated coefficients and their standard errors are the same as in H1, it follows that underrejection must be due to upward bias in the estimation of $\text{Cov}(\beta_d^{PPML}, \beta_d^{NLIV})$. These results on the Hausman tests are opposite in sign to previous findings concerning continuous endogenous regressors (Creel, 2004; Chmelarova, 2007), were Hausman contrast tests tend to overreject H_0. As for results on power, Table 2.1 displays rejection frequencies of the false null hypothesis under $\rho = 0.2$ (columns 4 to 6) and $\rho = 0.5$ (columns 7 to 9). The performance of H1 and H2 are practically indistinguishable. This implies that there might be very small or even no gains at all from implementing H2 instead of the more robust H1, even under an ideal DGP for H2.

Turning to the Wald tests, results are presented for tests based on the FIML estimates (FIML), two-stage method of moments estimates implemented via NLS (TSM NLS) and PPML (TSM PPML), as well as for the new test derived from the generalized residual inclusion (GRI) and the test based on the residual inclusion procedure (RI). The TSM tests are based on two-stage adjusted standard errors. For GRI and RI, results are presented separately for tests using regular standard errors and two-stage adjusted standard errors (GRI-TSA and RI-TSA). Thus, GRI and RI are tests which virtually can be implemented by the practitioner in a matter of seconds, while the two-stage adjustment might take more time as it generally requires a minimum of custom programming.

Considering the empirical size of the Wald tests with samples of 200 observations, most of them overreject the null hypothesis by 2 to 4 percentage points, with the exception of FIML and GRI-TSA, whose rejection frequencies are not significantly different from 5%.

With increasing sample size, the other tests also gradually tend to the nominal size. As the results make evident, using two-stage adjusted standard errors improves noticeably the empirical size of the GRI and RI tests in small to moderate samples, although the GRI-TSA standard errors seem to be a little bit too large leading to slight underrejection in some cases. The TSM NLS test is the only one to overreject clearly even with sample size 1'000. It also performs comparatively poorly with respect to power. As expected, FIML has the largest power in this setting where it is the efficient estimator, followed by the TSM PPML and GRI(-TSA) tests. The RI(-TSA) tests are comparable in power to the H1 Hausman test[12].

The DGP in Table 2.1 implied that $\text{Var}(\gamma_z z)/\text{Var}(\gamma_z z + \gamma_x x + v) = 0.25$, i.e., that the variance of the instrument determines one quarter of the total variance of the linear combination that determines d. Now, consider a change in instrument strength. By specifying a DGP which leaves $\gamma_z = \sqrt{0.5}$ as before, but with $\gamma_x = \sqrt{1.5}$, the fraction of the variance explained by the impact of the instrument, $\gamma_z z$, with respect to the whole systematic variance, $\text{Var}(\gamma_z z + \gamma_x x)$, falls from 0.5 to 0.25, while the systematic variance relative to the error variance, $\text{Var}(v)$, doubles. Taken together, the new instrument is weaker since $\text{Var}(\gamma_z z)/\text{Var}(\gamma_z z + \gamma_x x + v) \approx 0.167$. How does this change affect power and size of the tests? Comparing the columns with sample size 500 in Table 2.1 with columns labelled (2) in Table 2.2 gives an idea. While the Hausman and residual inclusion tests suffer severe power loss, TSM PPML and the generalized residual inclusion tests are barely affected. Figure 3.1 details the circumstance graphically by plotting the power functions of H1, TSM PPML, GRI-TSA and RI-TSA over the support of ρ for both DGPs. The difference in power grows with increasing absolute value of ρ and is over 20 percentage points at the extremes. The reason for this difference is that Hausman and residual inclusion tests rely

[12] Some authors prefer to use what is called size-corrected power to make comparisons across tests. Here, no size-corrected power is presented, since the question addressed is how these tests work in practice and which are useful under given characteristics of the data generating process.

only on the dependence between the instrument and the endogenous variable, which in this experiment was significantly weakened. Meanwhile, tests as TSM PPML and GRI seem to be able to compensate this loss with the increased variance of the systematic part which allows them to exploit more fully their functional form assumption.

The remaining columns in Table 2.2, labelled (1) and (3), show rejection frequencies of the null hypothesis for further instrument strength scenarios. Here $\text{Var}(\gamma_z z + \gamma_x x)$ is reset to unity as in Table 2.1, and only the fraction of it due to $\text{Var}(\gamma_z z)$ is modified to 0.25 (1) and 0.75 (3), inducing a weaker and stronger instrument, respectively. The results show that only GRI-TSA and RI-TSA reach appropriate size in the weak instrument case. In the scenario with the strong instrument, results are very similar to Table 2.1, with FIML capitalizing on its efficiency, followed by a more equalized middle field including H1, TSM PPML and their approximations GRI and RI. TSM NLS and H2 display markedly lower power, and H3 again falls prey to its strong underrejection problem.

Monte Carlo simulation studies always raise questions concerning the specificity of their results. To check that the presented results are not due to the particular choice of DGP, some sensitivity analysis has been conducted. First, orthogonal regressors are far from realistic in the social sciences. A further worry is the marginal distribution of the endogenous dummy, as in practice outcomes with 1 and 0 are often not balanced. Also, one may wonder if the tests are sensitive to a reduction of the effect of the dummy on the count variable. Finally, TSM and GRI are based on the null hypothesis $\theta = 0$, with $\theta = \sigma\rho$. Their positive performance could partly be due to the fact that in the shown DGP $\sigma = 1$ and so $\theta = \rho$. To address these concerns, separate simulations were carried out with $\text{Corr}(x, z) = 0.5$, $\text{E}(d|x, z) = 0.2$, $\beta_d = 0.1$ and $\sigma = \sqrt{2}$ (not reported). As it turns out, most results are by and large invariant to these alternatives. The exceptions are H1 and RI's reduced power when the regressors are correlated, as well as H1's when β_d is small. This latter finding is not surprising given that H1 is based on the contrast of estimates of β_d.

2.4.3 Identification by functional form

Having observed the performance of FIML, TSM PPML and GRI-TSA under reduced impact of the instrument (cf. Fig.1), a natural question is whether identification can be achieved by functional form alone, prescind from any instrument z. To this end, the DGP is specified as before, but setting $\gamma_z = 0$ and maintaining $\gamma_x = \sqrt{0.5}$. Results are shown in Table 2.3 in columns labelled (1) for sample sizes of 500 and 2'000 observations. The results prove to be rather discouraging, as both FIML and TSM PPML display empirical sizes that render the tests useless[13]. GRI-TSA's overrejection is not as pronounced, but the test lacks power in this setup. The exercise is repeated in columns (2) by strongly increasing the variance explained by the systematic part. To this end, γ_x is set to $\sqrt{2}$. However, little change is evident in the results for sample size 500. In the entries corresponding to the larger sample, on the other hand, some improvement is noticeable for TSM PPML and GRI-TSA, the latter's overrejection being only mild and showing increased power. Having empirical applications in mind, nevertheless, it seems that results from columns (1) represent a more realistic setting regarding instrument strength, so that the presence of an instrument in the DGP seems to be necessary for testing in finite samples.

2.4.4 Results under distributional misspecification

When specifying a parametric model, a natural concern relates to the robustness to distributional misspecification. In the context of count data, for instance, the overdispersion precluded from a Poisson distribution has been a major preoccupation which has led a portion of the empirical work to opt for the negative binomial regression model. Although under exogeneity the pseudo maximum likelihood properties of the Poisson model warrant

[13]Monfardini and Radice (2008) investigate exogeneity testing with no instruments in the bivariate probit model, which is related to the model under consideration through the bivariate normality assumption. The present results are in line with theirs, as they report high overrejection rates for Wald tests. They find likelihood ratio tests to have appropriate empirical size.

consistency of the estimator, in the model with endogenous binary variable presented here, FIML, TSM and GRI are inconsistent if ε and v are not normally distributed. Moreover, in general, Terza's (1998) FIML estimator yields inconsistent estimates whenever $f(y|x, d, \varepsilon)$ does not follow a Poisson distribution. However, Romeu and Vera-Hernandez (2005) show that in the case of the conditional distribution being Negative Binomial type I (NegBinI), the FIML estimator remains consistent, suggesting that so does the FIML test[14]. The first two columns in Table 2.4 illustrate the performance of selected tests under the baseline DGP from Table 2.1 but with the modification $y|x, d, \varepsilon \sim NegBinI$ with expectation λ as before, and variance 2λ. Only GRI-TSA displays correct size. FIML overrejects quite severely, while TSM PPML does less so, but has noticeably less power than in the baseline case. H1 underrejects and ranks as the least powerfull among the compared tests.

To assess sensitivity of test size to the crucial assumption of bivariate normality, a DGP is implemented where the errors (ε, v) are independent and follow marginal distributions different from the normal. The chosen distributions are the exp-Gamma(1,1) for ε, which combined with a Poisson distribution for y conditional on observables and ε, yields a Neg-BinI distribution for y conditional on observables only; and a logistic distribution for v, scaled as to have unit variance, which gives a logit model for d. It might be argued that these modifications represent rather moderate departures from the distributional assumptions. However, there are at least two reasons for considering such a scenario. First, as mentioned before, there is a large body of empirical literature that uses NegBin and logit models, which consequently must imply either that there exists a large number of real-world problems where assuming negative binomial and logit processes is sensible, or that said literature's distributional assumptions are wrong. The former reason might find wider approval. Second, if the researcher has a strong belief in some form of significant departure

[14]Corollary 1 in Romeu and Vera-Hernandez (2005) establishes consistency of $(\hat{\beta}, \hat{\beta}_d)$ excluding the constant element, which is shifted. The estimate $\hat{\rho}$ is inconsistent for ρ but equals 0 whenever ρ does, securing consistency of the exogeneity test.

from normality of the errors which goes beyond exp-Gamma or logit, she might as well opt to model this explicitly. Further, one might be interested in the performance of the tests under mild misspecification, since tests that do not conform to one's expectations even under these circumstances might as well be regarded as useless in view of the inherent uncertainty faced with respect to the 'true' data generating process. In other words, rather than her assumptions coinciding exactly with reality, all the applied econometrician might hope is that her assumptions approximate the underlying data generating process reasonably well.

Setting these concerns apart and considering the results of this analysis as shown in the third column in Table 2.4, the tests do present some minor size distortions, with H1 and GRI-TSA underrejecting, and TSM PPML and RI-TSA overrejecting H_0. FIML's overrejection is more substantial. In order to analyze empirical power of the tests under non-normal marginals, dependence between the errors is induced by random sampling from copula functions. Columns 4 and 5 in Table 2.4 show rejection frequencies of the null hypothesis of exogeneity when the errors' joint distribution is generated from a bivariate Gaussian copula with exp-Gamma and logistic marginals, with dependence parameter θ^{GC} equal to 0.2 and 0.5, respectively. Note that θ^{GC}, although having the same domain, is not a correlation coefficient as in the bivariate normal distribution, and thus comparisons to other tables are not valid. However, both columns reproduce the familiar pattern of the more parametric tests outperforming the supposedly more robust ones. Also, RI-TSA, which displayed power comparable to H1, clearly surpasses H1 in this setting. The last two columns in Table 2.4 contain results obtained by letting the joint distribution of the errors be determined by a Frank copula with the same non-normal marginals as before. The Frank copula induces positive dependence between the variables through the parameter $\theta^{FC} \in (0, \infty)$, with independence resulting as a special case when $\theta^{FC} = 0$. The parameter is set to 1 in the sixth column and to 10 in the seventh column in Table 2.4. While for the weaker dependence power between the tests is rather similar, differences

are considerably more pronounced for the case of stronger dependence. The ranking of the tests is almost the same as with the Gaussian copula, except for FIML falling back to third place. On the whole, these results seem to indicate that the tests relying on the bivariate normality assumption might perform equally well in non-normal settings as the other tests. Furthermore, GRI-TSA's actual type I error seems never to be larger than the level determined by the nominal size.

2.4.5 Exogeneity tests as pretests: A cautionary note

By far the most common use of tests for exogeneity is probably as pretests in order to choose between estimates. If a test rejects exogeneity, then estimates are obtained from an estimator that is consistent under endogeneity; while if the tests fails to reject the exogeneity hypothesis, estimates can be calculated from an estimator that is efficient under exogeneity, although inconsistent if the true DGP entails endogeneity. Thus, inference about a parameter of interest is conditional on the outcome of the exogeneity pretest.

The pretests or first stage tests to be considered are the exogeneity tests discussed so far, H1, FIML, TSM PPML, GRI-TSA and RI-TSA. If the pretest fails to reject the null hypothesis, the model is estimated by Poisson MLE and a (second stage) two-tailed t-test with null hypothesis $H_0 : \beta_d = 1$ is conducted. Given rejection of exogeneity in the first stage test, the second stage test of $H_0 : \beta_d = 1$ is performed with NLIV estimates if the pretest was either H1 or RI-TSA. For TSM PPML and GRI-TSA pretests, second stage tests are calculated with TSM PPML estimates, while FIML pretests use FIML estimates in the second stage[15]. In the DGP, the true β_d is left at 1 throughout all simulations, so that empirical rejection frequencies measure the finite sample size of the second stage test.

Inspection of the results displayed in Table 2.5 suggests that the use of pretests for exogeneity leads to severe size distortions unless $\rho = 0$. Moreover, the overrejection is increasing over the range of ρ shown in the table, except for FIML. The reason for this

[15]Second stage tests do not use RI-TSA and GRI-TSA estimates as these are inconsistent unless $\rho = 0$.

is that for weaker levels of correlation, the weak power of the pretests leads to second stage tests being performed with Poisson ML estimates whose bias for low ρ is sufficiently small as to not always reject H_0. Loosely speaking, as ρ increases, the bias in β_d increases faster than the power of the pretests, leading to higher rejection frequencies for all tests. Eventually, all second stage tests' overrejection lowers, but except for FIML the turning point is after $\rho = 0.5$.

It is clear from the estimated rejection frequencies which are nowhere near the nominal size, that inference on structural parameters after pretesting in this model is likely to lead to false results and should thus be avoided. It should be stressed, however, that the pernicious effect of pretesting is due to interpreting the failure to reject exogeneity as that the variable in question is exogenous (*absence* of endogeneity). Obviously, exogeneity tests can be used to provide empirical evidence of the *presence* of endogeneity. This can be important in its own right, as for putting theories to test, and it can also provide ex-post empirical confirmation for a-priori concerns about potential endogeneity.

2.5 Conclusions

In this article some tests for exogeneity of a binary variable in count data regression models, including the new GRI test, were examined for their finite sample properties through Monte Carlo simulations. The behavior of the tests under correct distributional specification was analyzed subjecting them to different sample sizes and levels of instrument strength. Test performances under data generating processes with no instrumental variables were reported, as well as under distributional misspecification. Finally, the use of these tests as pretests was assessed. Based on the results of the Monte Carlo experiments, a number of conclusions can be drawn which might provide some guidance for empirical practice.

The Hausman test which contrasts Poisson ML and NLIV estimates (H1) performs better than the other more refined versions based on Poisson-log-normal estimates (H2) or

on estimation of the covariance between estimates (H3). Tests based on residual inclusion (RI) represent a very easy to implement alternative to H1, which in most scenarios display power comparable to H1, while outperforming Hausman contrast tests with respect to empirical size.

The other more parametric Wald tests which are based on the bivariate normality assumption generally present higher power than the Hausman tests, even in settings where they misspecify the DGP. The FIML test generally achieves the highest power of the tests. The more robust approximation to FIML, TSM, works well when it is implemented through PPML instead of NLS, achieving power almost as high as FIML. The first order approximation to FIML, generalized residual inclusion (GRI), exhibits slightly lower power than TSM PPML, but still performs favorably compared to H1.

On the whole, therefore, these results suggest that using the simpler RI and GRI tests comes at virtually no cost in terms of test performance. Using two-stage adjusted standard errors noticeably improves the empirical size of the tests in smaller samples. Moreover, these tests show the best performances of all tests in the smallest samples and under the weakest instrument strength levels that were used in the simulations.

Two caveats have to be considered when testing for exogeneity. The first relates to the absence of exclusion restrictions in the DGP. Only with large samples and a very strong instrument does GRI-TSA come close to the nominal test size, the other tests perform worse. This suggests that there is little hope to test for endogeneity in practice if the structural model does not include any instruments.

The second issue concerns the use of these tests as pretests. In line with Guggenberger's (2008) finding of severe size distortions conditional on Hausman pretests in the classical linear model, large overrejection rates render pretesting futile in the present count data model. The higher power of the Wald pretests clearly is not enough to result in acceptable second stage sizes. Therefore, practitioners are well advised to avoid using these tests as pretests. However, given that theoretical concerns about endogeneity have led a researcher

to implement an estimation procedure that accounts for this, endogeneity tests can be used to obtain ex-post empirical evidence of these concerns having been justified.

References

Angrist, Joshua (2001), Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors: Simple Strategies for Empirical Practice, *Journal of Business and Economic Statistics*, **19**(1), 2-16.

Benavides-Velasco, Carlos A. and Cristina Quintana-Garcia (2008), Innovative competence, exploration and exploitation: The influence of technological diversification, *Research Policy*, **37**(3), 492-507.

Chesher, Andrew (1991), The effect of measurement error, *Biometrika*, **78**(3), 451-462.

Chmelarova, Viera (2007), The Hausman test, and some alternatives, with heteroscedastic data, Dissertation, Department of Economics, Louisiana State University.

Creel, Michael (2004), Modified Hausman tests for inefficient estimators, *Applied Economics*, **36**, 2373-2376.

Dagenais, Marcel G. (1999), Inconsistency of a proposed nonlinear instrumental variables estimator for probit and logit models with endogenous regressors, *Economics Letters*, **63**(1), 19-21.

Degeorge, Francois, Francois Derrien and Kent L. Womack (2007), Analyst Hype in IPOs: Explaining the Popularity of Bookbuilding, *Review of Financial Studies*, **20**(4), 1021-1058.

Egger, Peter, Mario Larch, Kevin E. Staub and Rainer Winkelmann (2009), The trade effects of endogenous preferential trade agreements, Manuscript.

Gourieroux, Christian S. and Michael Visser (1997), A count data model with unobserved heterogeneity, *Journal of Econometrics*, **79**(2), 247-268.

Greene, William H. (1995), Sample selection in the Poisson regression model, Working Paper, Department of Economics, Stern School of Business, New York University.

Greene, William H. (1998), Sample selection in credit-scoring models, *Japan and the World Economy*, **10**(3), 299-316.

Grogger, Jeffrey (1990), A simple test for exogeneity in probit, logit and Poisson regression models, *Economics Letters*, **33**(4), 329-332.

Guggenberger, Patrik (2008), The impact of a Hausman pretest on the size of hypothesis tests, Working Paper (first version: 2006), Department of Economics, University of California, Los Angeles.

Hausman, Jerry A. (1978), Specification tests in econometrics, *Econometrica*, **46**, 1251-1271.

Kenkel Donald S. and Joseph V. Terza (2001), The effect of physician advice on alcohol consumption: Count regression with an endogenous treatment effect, *Journal of Applied Econometrics*, **16**, 165-184.

Kozumi, Hideo (2002), A bayesian analysis of endogenous switching models for count data, *Journal of the Japanese Statistical Society*, **32**(3), 141-154.

Masuhara, Hiroaki (2008), Semi-nonparametric count data estimation with an endogenous binary variable, *Economics Bulletin*, **42**(3), 1-13.

Miranda, Alfonso (2004), FIML estimation of an endogenous switching model for count data, *Stata Journal*, **4**(1), 40-49.

Monfardini, Chiara and Rosalba Radice (2008), Testing exogeneity in the bivariate Probit model: A Monte Carlo study, *Oxford Bulletin of Economics and Statistics*, **70**(2), 271-282.

Mullahy, John (1997), Instrumental variable estimation of count data models: Applications to models of cigarette smoking behavior, *Review of Economics and Statistics*, **79**, 586-593.

Nichols, Austin (2007), IVPOIS: Stata module to estimate an instrumental variables Poisson regression via GMM, available online at http://ideas.repec.org/c/boc/bocode/s456890.html

Orme, Chris D. (2001), Two-step inference in dynamic non-linear panel data models, Manuscript, School of Economic Studies, University of Manchester.

Parrado, Emilio A., Chenoa A. Flippen and Chris McQuiston (2005), Migration and Relationship Power Among Mexican Women, *Demography*, **42**, 347-372.

Romeu, Andres and Marco Vera-Hernandez (2005), Counts with an endogenous binary regressor: A series expansion approach, *Econometrics Journal*, **8**, 1-22.

Santos Silva, Joao M.C. and Silvana Tenreyro (2006), The log of gravity, *Review of Economics and Statistics*, **88**(4), 641-658.

Terza, Joseph V. (1998), Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects, *Journal of Econometrics*, **84**(1), 129-154.

Terza, Joseph V. (2006), Estimation of policy effects using parametric nonlinear models: a contextual critique of the generalized method of moments, *Health Services and Outcomes Research Methodology*, **6**(3-4), 177-198.

Terza, Joseph V., Anirban Basu and Paul J. Rathouz (2008), Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling, *Journal of Health Economics*, **27**, 531-543.

Weesie, Jeroen (1999), Seemingly unrelated estimation and the cluster-adjusted sandwich estimator, *Stata Technical Bulletin*, **52**, 34-47.

Windmeijer, Frank A.G. and João M.C. Santos Silva (1997), Endogeneity in count data models: An application to demand for health care, *Journal of Applied Econometrics*, **12**(3), 281-294.

Winkelmann, Rainer (2008), *Econometric Analysis of Count Data*, fifth edition, Berlin: Springer.

Wooldridge, Jeffrey M. (1997), "Quasi-Likelihood Methods for Count Data", in M. Hashem Pesaran and Peter Schmidt (eds.), *Handbook of Applied Econometrics, Volume II: Microeconomics*, Massachusets, USA/Oxford, UK: Blackwell Publishers, 352-406.

Table 2.1: Rejection frequencies of tests for exogeneity - The effect of sample size

	$\rho = 0$			$\rho = 0.20$			$\rho = 0.50$		
Sample size:	200	500	1000	200	500	1000	200	500	1000
Hausman contrast tests									
H1	0.0365	0.0459	0.0517	0.0672	0.1168	0.2019	0.1796	0.4423	0.7451
H2	0.0287	0.0371	0.0432	0.0583	0.1050	0.1798	0.1708	0.4223	0.7239
H3	0.0038	0.0060	0.0084	0.0097	0.0265	0.0534	0.0363	0.1902	0.4788
Wald tests									
FIML	0.0540	0.0635	0.0640	0.0670	0.1600	0.2750	0.2070	0.6605	0.9160
TSM NLS	0.0893	0.0728	0.0627	0.0668	0.0638	0.0799	0.0790	0.1997	0.4376
TSM PPML	0.0739	0.0616	0.0561	0.0766	0.1079	0.1806	0.2046	0.4616	0.7620
GRI	0.0750	0.0603	0.0573	0.1047	0.1309	0.1958	0.2798	0.4971	0.7570
RI	0.0814	0.0605	0.0554	0.1060	0.1240	0.1706	0.2445	0.3964	0.5963
GRI-TSA	0.0509	0.0441	0.0420	0.0748	0.0999	0.1578	0.2188	0.4287	0.7043
RI-TSA	0.0711	0.0566	0.0535	0.0945	0.1192	0.1667	0.2272	0.3863	0.5931

Notes: Number of replications = 10'000 (FIML: 2'000 replications). Nominal test size = 0.05.

Table 2.2: Rejection frequencies of tests for exogeneity - The effect of instrument strength

	$\rho = 0$			$\rho = 0.20$			$\rho = 0.50$		
IV strength:	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Hausman contrast tests									
H1	0.0283	0.0449	0.0584	0.0639	0.0936	0.1541	0.2206	0.3042	0.5954
H2	0.0248	0.0375	0.0413	0.0602	0.0826	0.1286	0.2194	0.2995	0.5560
H3	0.0072	0.0086	0.0040	0.0222	0.0277	0.0217	0.1091	0.1325	0.2223
Wald tests									
FIML	0.0820	0.0620	0.0555	0.1305	0.1525	0.2015	0.4935	0.6095	0.7710
TSM NLS	0.0819	0.0862	0.0695	0.0622	0.0744	0.0799	0.1100	0.2080	0.2905
TSM PPML	0.0666	0.0683	0.0566	0.0960	0.1071	0.1256	0.3111	0.4382	0.5841
GRI	0.0629	0.0640	0.0586	0.1009	0.1206	0.1528	0.3408	0.4543	0.6055
RI	0.0617	0.0633	0.0594	0.0951	0.0980	0.1494	0.2496	0.2665	0.5174
GRI-TSA	0.0451	0.0484	0.0419	0.0779	0.0964	0.1168	0.2835	0.3984	0.5460
RI-TSA	0.0533	0.0581	0.0577	0.0848	0.0908	0.1468	0.2315	0.2544	0.5130

Notes: Number of replications = 10'000 (FIML: 2'000 replications). Nominal test size = 0.05. IV-strength as detailed in text or Table 6.

Table 2.3: Rejection frequencies of tests for exogeneity - Identification by functional form

	(1)			(2)		
	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
N=500						
FIML	0.1565	0.1750	0.2640	0.1375	0.1775	0.4155
TSM PPML	0.1783	0.1960	0.2473	0.1179	0.1181	0.2585
GRI-TSA	0.0729	0.0677	0.0860	0.0812	0.0838	0.2001
N=2000						
FIML	0.2340	0.2950	0.5700	0.1630	0.3490	0.8640
TSM PPML	0.1643	0.1700	0.2990	0.0780	0.1438	0.6538
GRI-TSA	0.0772	0.0714	0.1327	0.0610	0.1123	0.5546

Notes: Number of replications = 10'000 (FIML: 2'000 replications for N=500, 1'000 replications for N=2'000). Nominal test size = 0.05. IV-strength of columns (1) and (2) as detailed in text or Table 6.

Table 2.4: Rejection frequencies of tests for exogeneity - Sensitivity to distributional assumptions

	NegBin I		Gaussian copula			Frank copula	
	$\rho = 0$	$\rho = 0.5$	$\theta = 0$	$\theta^{GC} = 0.2$	$\theta^{GC} = 0.5$	$\theta^{FC} = 1$	$\theta^{FC} = 10$
H1	0.0382	0.3321	0.0415	0.0647	0.2930	0.0856	0.5833
FIML	0.1035	0.5970	0.0470	0.1445	0.5870	0.0820	0.7245
TSM PPML	0.0608	0.3747	0.0596	0.1546	0.5859	0.1008	0.9197
GRI-TSA	0.0451	0.6243	0.0400	0.1186	0.5359	0.0817	0.8317
RI-TSA	0.0582	0.5248	0.0582	0.1139	0.4404	0.0917	0.7064

Notes: Number of replications = 10'000 (FIML: 2'000 replications). Nominal test size = 0.05. Sample size = 500. IV-strength as detailed in text or Table 6.

Table 2.5: Empirical size of second stage tests of $\beta_d = 1$ using pretests for exogeneity

	(1)			(2)		
	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
H1	0.0383	0.3596	0.5507	0.0409	0.3148	0.3960
FIML	0.0540	0.3565	0.3365	0.0520	0.3055	0.2270
TSM PPML	0.0516	0.3681	0.5327	0.0495	0.3326	0.4082
GRI-TSA	0.0493	0.3584	0.4981	0.0471	0.3219	0.3877
RI-TSA	0.0366	0.3578	0.6069	0.0382	0.3187	0.4767

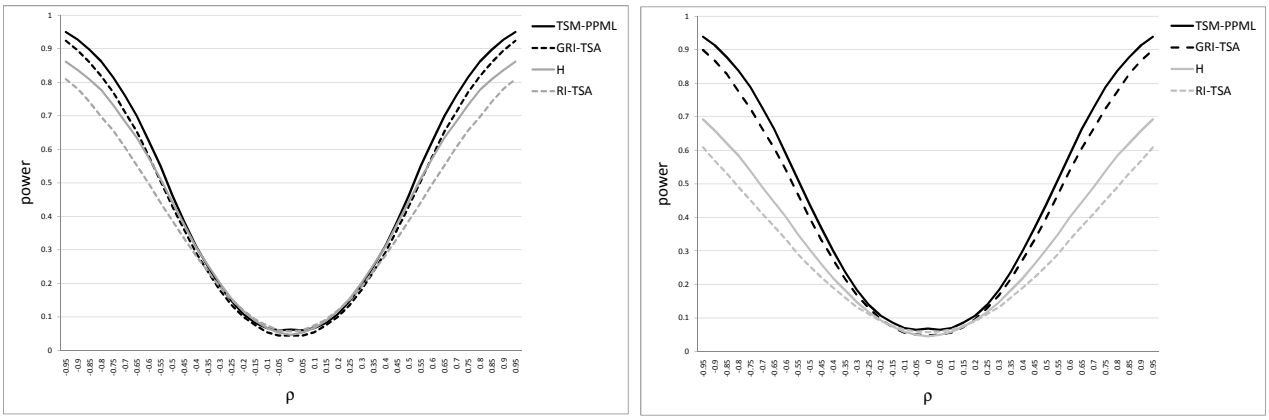
Notes: Number of replications = 10'000 (FIML: 2'000 replications). Nominal test size = 0.05. Sample size = 500. IV-strength of columns (1) and (2) as detailed in text or Table 6.

Table 2.6: Details on the DGP of Monte Carlo simulations

Table	Columns	Distribution of $y\|x,d,\varepsilon$	Distribution of (ε, v)	Reduced form parameters (γ_x, γ_z)
1	all	$Poisson(\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{0.50}, \sqrt{0.50})$
2	(1)	$Poisson(\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{0.75}, \sqrt{0.25})$
	(2)	$Poisson(\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{1.50}, \sqrt{0.50})$
	(3)	$Poisson(\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{0.25}, \sqrt{0.75})$
3	(1)	$Poisson(\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{0.50}, 0.00)$
	(2)	$Poisson(\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{2.00}, 0.00)$
4	(1), (2)	$NegBin(\lambda,\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{0.50}, \sqrt{0.50})$
	(3)	$Poisson(\lambda)$	$\varepsilon \sim expGamma(1,1)$, $v \sim Logistic(0, 3/\pi)$	$(\sqrt{0.50}, \sqrt{0.50})$
	(4), (5)	$Poisson(\lambda)$	Gaussian copula*	$(\sqrt{0.50}, \sqrt{0.50})$
	(6), (7)	$Poisson(\lambda)$	Frank copula*	$(\sqrt{0.50}, \sqrt{0.50})$
5	(1)	$Poisson(\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{0.50}, \sqrt{0.50})$
	(2)	$Poisson(\lambda)$	$BVN(0,0,1,1,\rho)$	$(\sqrt{0.25}, \sqrt{0.75})$

* Marginal distributions of the copulae: $\varepsilon \sim expGamma(1,1)$, $v \sim Logistic(0, 3/\pi)$.

Figure 2.1: Empirical power of tests for exogeneity



Notes: Sample size = 500. Nominal test size = 0.05. Reduced form parameters: Left panel $(\gamma_x, \gamma_z) = (\sqrt{0.5}, \sqrt{0.5})$; right panel $(\gamma_x, \gamma_z) = (\sqrt{1.5}, \sqrt{0.5})$. Graphs based on 20 points $\rho = 0, 0.05, 0.10, ..., 0.95$. Values for negative ρ mirrored symmetrically from corresponding positive points. Each point obtained from 10'000 replications.

Chapter 3

Quasi-likelihood estimation of zero-inflated count models

This chapter is joint work with Rainer Winkelmann. It is a revised version of Working Paper No. 0908, *SOI Working Paper Series*, Department of Economics, University of Zurich.

Acknowledgements: We thank participants of the 2009 Engelberg Workshop in Labor Economics, the 2010 New Zealand Econometric Study Group Meeting in Auckland and the 2010 Annual Meeting of the Swiss Society of Economics and Statistics in Fribourg for valuable comments.

3.1 Introduction

The Poisson regression model is the benchmark model for regressions with count dependent variables. The so-called problem of "excess zeros", however, plagues a majority of count data applications in the social sciences, as the proportion of observations with zero counts in the sample is often much larger than that predicted by the Poisson model. The conventional wisdom of the pertinent literature is that with "excess zeros", Poisson regression should be abandoned in favor of modified count data models which are capable of taking into account the extra zeros explicitly. By far the most popular of these models are zero-inflated (ZI) count models (Mullahy, 1986, Lambert, 1992) such as the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models. Applications often feature a separate logit or probit model for the excess zeros. Recent examples include job interviews (List, 2001), work absences (Campolieti, 2002), job changes (Heitmueller, 2004), lateness (Clark et al., 2005), patent applications (Stephan et al., 2007), cigarette consumption (Sheu et al., 2004), theatre attendance (Ateca-Amestoy, 2008), biking trips (Zahran et al., 2008) and firm FDI (Ho et al., 2009).

This article is concerned with estimation of regression parameters for count data when there are excess zeros but the exact distribution of the counts is uncertain. Using the framework for maximum likelihood estimation of misspecified models by Gourieroux, Monfort and Trognon (1984a), it can be shown that ZIP and ZINB are inconsistent unless correctly specified. As an easily implementable alternative, we propose a new Poisson Quasi-Likelihood (PQL) estimator. This estimator can accommodate a logit regression part for the excess zeros. In the case of constant zero-inflation, it can be estimated with standard Poisson software. Otherwise, a modification is required. The new PQL estimator is robust to misspecification, as it estimates the regression parameters consistently regardless of the true distribution for the counts. A series of small Monte Carlo experiments shows that PQL estimation is free of bias in moderate samples, whereas the ZIP and ZINB can have sizeable biases.

The next section presents the standard zero-inflated models and discusses some of their limitations. The new PQL estimators for ZI count models are discussed in Section 3.3. Section 3.4 presents Monte Carlo simulation results comparing PQL to the ML estimators. Section 3.5 illustrates the new PQL estimator with logit zero-inflation for modeling the frequency of doctor visits. Section 3.6 concludes.

3.2 Modeling "excess zeros"

Zero-inflated count data models have generic probability function (pf)

$$f(y) = \begin{cases} \pi + (1-\pi)f^*(0) & \text{for } y = 0 \\ (1-\pi)f^*(y) & \text{if } y = 1, 2, 3, ... \end{cases} \tag{3.1}$$

where y is a count-valued random variable. The function $f^*(\cdot)$, called the parent model, is a standard count pf, and $\pi \in [0, 1]$ is a zero-inflation parameter which allows for additional zeros. Zero-inflated models introduce a distinction between so-called 'structural' or 'strategic' zeros that are due to the inflation part, and 'incidental' zeros stemming from the count distribution part of the model. For instance, considering job mobility, a person might not have changed job in a given time period because she is not looking for a new one (structural zero) or because despite searching she has not found another job (incidental zero). Similarly, when consider demand for, say, movies (or any other consumer item), a zero may indicate that the individual either never goes to the movies, or else goes occasionally but did not do so in the reference period.

If $\pi = 0$, the ZI pf $f(\cdot)$ reduces to the parent model. The two most common choices for $f^*(\cdot)$ are Poisson,

$$f^{\mathrm{P}}(y; \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \ , \ \lambda > 0$$

and negative binomial,

$$f^{\mathrm{NB}}(y; \lambda) = \frac{\Gamma(\gamma + y)}{\Gamma(\gamma)\Gamma(y+1)} \left(\frac{\gamma}{\lambda+\gamma}\right)^{\gamma} \left(\frac{\lambda}{\lambda+\gamma}\right)^{y} \ , \lambda > 0 \, , \, \gamma > 0$$

Both models' expectation is equal to λ, which also gives the variance in the Poisson case. The variance in the negative binomial model is $\lambda + \gamma^{-1}\lambda^2$. Let x be a vector of explanatory variables including a constant. In a regression context, it is customary to specify the mean parameter λ as an exponential function of x

$$\lambda = \exp(x'_{it}\beta) \tag{3.2}$$

where β is a parameter vector conformable to x. The linear predictor can include polynomial expansions that approximate a non-linear function to any desired degree of precision. Many applications generalize the model defined by equations (3.1) and (3.2) to allow for non-constant zero-inflation by specifying π as a function of covariates, for example a logit model (as in Lambert, 1992):

$$\pi = \frac{\exp(z'\delta)}{1 + \exp(z'\delta)} \tag{3.3}$$

z can be identical to x, overlap, or be completely distinct. The parameters of the ZIP and ZINB models are usually estimated by Maximum Likelihood (ML). The log-likelihood function for the ZIP model for a sample of n independent observation tuples (y_i, x_i, z_i) is

$$\begin{aligned} \ln l^{ZIP} &= \sum_{i=1}^{n} \mathbb{1}(y_i = 0) \ln[\exp(z'_i\delta) + \exp(-\exp(x'_i\beta))] \\ &\quad + \mathbb{1}(y_i > 0)[-\exp(x'_i\beta) + y_i x_i \beta] - \ln(1 + \exp(z'_i\delta)) \end{aligned} \tag{3.4}$$

Since these models have a finite mixture structure, maximization of the log-likelihood function can employ the EM algorithm, although direct maximization using Newton-Raphson is possible as well. If the model – consisting of equations (3.1), (3.2), (3.3) and $f^*(\cdot)$ – is correctly specified, ML theory ensures that these estimators are consistent and asymptotically efficient, provided they exist (Cameron and Trivedi, 1998; Winkelmann, 2008).

Sometimes, convergence problems are reported. The potential instability of ML estimation of ZI count models in the presence of outliers has been noted by Hall and Sheng (2010) who also explore an alternative robust estimation technique. Failure to converge may also

be symptomatic for a non-existing ML estimator. Consider again the log-likelihood function (3.4) and assume that one of the regressors z_k is a partially discrete variable such that

$$z_k \quad \begin{cases} \geq 0 & \text{for } y > 0 \\ = 0 & \text{for } y = 0 \end{cases}$$

Then, the first-order condition of the ZIP for the associated parameter δ_k is

$$\sum_{y_i>0} -\frac{\exp(z_i'\delta)}{1+\exp(z_i'\delta)} z_{ik} = 0$$

which has no solution so that the ZIP estimator does not exist. A similar argument applies to the ZINB. This is a "perfect prediction" problem common to non-linear binary choice models (e.g., Albert and Anderson, 1984). Although the prospect of perfect prediction is not the main reason speaking for the PQL estimator we discuss next, it is noteworthy that the latter does not suffer from the same particular deficiency.

3.3 PQL estimation of zero-inflated models

The development of our new model is motivated by the lack of robustness of conventional ZIP and ZINB estimators if the underlying distribution assumptions are violated. Suppose that the key objects of interest are the CEF $\mathrm{E}(y|x)$ and its semi-elasticities $\partial[\mathrm{E}(y|x)/\mathrm{E}(y|x)]/\partial x_k$. With λ defined as in (3.2) and constant ZI parameter π, the conditional expectation function of the zero inflated model is given by

$$\mathrm{E}(y|x) = (1-\pi)\lambda = \exp(\ln(1-\pi) + x_{it}'\beta) \tag{3.5}$$

The only difference to the CEF of the parent model is a constant shifted by $\ln(1-\pi)$. Hence it is not possible to separately identify π and the constant in the parent model, say β_0. This is of secondary importance in most applied work, since only knowledge of the overall constant, $\tilde{\beta}_0 = \ln(1-\pi) + \beta_0$, is needed in order to compute the CEF and semi-elasticities.

The latter are given by

$$\frac{\partial \mathrm{E}(y|x)/\mathrm{E}(y|x)}{\partial x_k} = \beta_k,$$

and therefore identical to those of the parent model.

In order to consistently estimate the parameters of the CEF and the resulting semi-elasticities, any moment based estimator can be used, for example NLS. In particular, estimation based on the Poisson regression with CEF (3.5) is consistent as well. This follows since the Poisson distribution is a member of the linear exponential family (LEF), which is the class of distributions with pf of the form

$$f^{\mathrm{LEF}}(y|\mu_x) = \exp\{a(\mu_x) + b(y) + c(\mu_x)y\}, \quad \text{where } \mu_x = \mu(x;\beta) = \mathrm{E}(y|x),$$

for $a(\mu_x) = -\mu_x$, $b(y) = -\ln(y!)$ and $c(\mu_x) = \ln(\mu_x)$. LEFs have the property that the score function can be written as

$$\frac{\partial \log f(y|x)}{\partial \beta} = (y - \mu_x)h(x) \tag{3.6}$$

where $h(x) = [dc(\mu_x)/d\mu_x][\partial \mu_x/\partial x]$. Suppose the true model is $g_0(y|x) \neq f(y|x)$ but $\mathrm{E}_0(y|x) = \mu_x$ for some value β_0. Thus, the CEF is correctly specified. In this case, the expectation of (5.8) at the true density is zero, even though the model is misspecified, since the CEF residual $y - \mathrm{E}(y|x)$ is independent of x, and thus has zero covariance with any function $h(x)$. As the empirical score converges to the expected score by the law of large numbers, the solution to the ML first order conditions converges in probability to the true CEF parameters as long as the CEF is correctly specified and (first-order) identified. This holds regardless of misspecification of higher conditional moment functions as long as a LEF distribution such as the Poisson is used for constructing the quasi-likelihood function (White, 1982; Gourieroux, Monfort and Trognon, 1984a, 1984b). Even though the data are zero-inflated, a simple Poisson regression gives valid estimates of the objects of interest as long as the CEF is correctly specified. Valid standard errors require an adjustment to the covariance matrix.

Unlike the Poisson model, the ZIP and ZINB models are not LEF members. Indeed, the log-probability function of a ZIP variable is

$$\ln f^{ZIP}(y;\lambda,\pi) = \mathbb{1}(y=0)[\ln(\pi+(1-\pi)\exp(-\lambda))] + (1-\mathbb{1}(y=0))[\ln(1-\pi)-\lambda+y\ln\lambda-\ln(y!)]$$

which cannot be written as $a(\mu)+b(y)+c(\mu)y$ with $\mu=(1-\pi)\lambda$, as there is no way of isolating an additive component that is linear in y —i.e. $c(\mu)y$— due to the (nonlinearity of the) indicator function $\mathbb{1}(y=0)$. An analog argument holds for the ZINB log-probability and, in fact, for any ZI model generated according to (3.1). From Theorem 2 in Gourieroux, Monfort and Trognon (1984a), LEF membership is a necessary condition for consistency of a quasi likelihood estimator. Consequently, misspecification of higher conditional moments will in general lead to asymptotic bias in these models.

Next consider the case of non-constant zero-inflation, with π specified as a parametric function of covariates. Here, the discussion is limited to logit zero-inflation (see equation (3.3)) as it is most widely represented in the existing literature. The conventional way in which the literature has opted to estimate these models is by modifying the constant ZI models' log-likelihood function to accommodate the function of the logit model. Thus, ZIP and ZINB estimators for this model are obtained by maximizing the corresponding log-likelihood functions (equation (3.4) for the ZIP) with respect to $\theta=(\beta,\delta)$ for ZIP and with respect to $\theta=(\beta,\delta,\gamma)$ for ZINB. If the assumed model is equal to the underlying data generating process, these estimators are consistent and asymptotically efficient. Under misspecification, they are inconsistent.

Again, a robust estimator can be obtained by using only information on the CEF. The CEF of the model with logit zero-inflation is given by

$$\mathrm{E}(y|x,z) = (1-\pi)\lambda = \frac{\exp(x_{it}'\beta)}{1+\exp(z'\delta)} \tag{3.7}$$

The PQL estimator for the model with non-constant zero-inflation is thus obtained by

maximizing

$$ql(\beta, \delta) = \sum_{i=1}^{n} y_i \ln \tilde{\lambda}_i - \tilde{\lambda}_i \tag{3.8}$$

where $\tilde{\lambda}_i = \exp(x_i'\beta)/(1 + \exp(z_i'\delta))$. Maximizing of (3.8) using the Newton-Raphson or related algorithms is relatively straightforward (Stata code is available on request). The first-order conditions are

$$\frac{\partial ql(\beta, \delta)}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - \frac{\exp(x_i'\beta)}{1 + \exp(z_i'\delta)} \right) x_i = 0$$

and

$$\frac{\partial ql(\beta, \delta)}{\partial \delta} = \sum_{i=1}^{n} \left(\frac{\exp(x_i'\beta + z_i'\delta)}{(1 + \exp(z_i'\delta))^2} - \frac{\exp(z_i'\delta)}{1 + \exp(z_i'\delta)} y_i \right) z_i = 0$$

This new estimator for zero-inflated count data is formally identical to the Poisson-logit model for underreported counts discussed by Winkelmann and Zimmermann (1993) (see also Papadopoulos and Santos Silva, 2008). It is consistent even if the true data generating process is not Poisson distributed - as is the case by definition with excess zeros. Of course, there are other estimators that can be used to estimate the parameters of interest consistently based on the appropriate CEF specification, such as nonlinear least squares (NLS) and various moment-based estimators. Among those, PQL has the appeal of simplicity, as its first order conditions are plain orthogonality conditions between residuals and regressors. Other estimators introduce weighting schemes the choice of which can affect efficiency. Exploiting these potential efficiency gains requires making additional assumptions on higher order moments.

While the PQL approach advocated in this article has some well defined advantages over ZIP and ZINB modeling, it is not a panacea. First, estimation for the model with non-constant zero-inflation is feasible only if some constraints on the relationship between x and z hold. A sufficient condition for identification of β and δ is the existence of an element in z that is excluded from x (Papadopoulos and Santos Silva, 2008). Second, one might want to predict probabilities of certain events or elasticities of such probabilities to specific

regressors. Using the PQL estimates with the Poisson pf for this purpose is inappropriate as more structure is needed. Third, PQL can be less efficient than the ZI estimator if the zero-inflated model is correctly specified.

3.4 Monte Carlo evidence

A Monte Carlo study has been conducted to assess the small sample properties of the various approaches for estimating models with extra zeros, and to compare their relative efficiency. Furthermore, the Monte Carlo study investigates biases arising from distributional misspecifications in ZIP and ZINB, illustrating the robustness of PQL in such cases.

3.4.1 Simulation design

Monte Carlo experiments are conducted for three distinct setups. For all three, the basic design of the experiment is as follows. The count dependent variable y is specified as

$$y = \begin{cases} 0 & \text{with probability } \pi \\ y^* & \text{with probability } 1-\pi \end{cases}$$

where $y^* \sim \text{Poisson}(\lambda)$, and λ and π are given by

$$\lambda = \exp(\alpha + \beta x + v), \quad \pi = \frac{\exp(\delta_0 + \delta_1 z)}{1 + \exp(\delta_0 + \delta_1 z)}$$

with the scalar regressors x and z being jointly normally distributed and having a correlation of 50%, $(x, z) \sim \text{BVN}(0, 0, 1, 1, 0.5)$. The primary focus is on estimation of β, which is set to 1. The parameter α is set to 0.5. The parameters of the logit zero-inflation (δ_0, δ_1) are varied to obtain different data generating processes. By letting $\delta_1 = 0$, a constant zero-inflation model is obtained. $\delta_1 = 1$ gives a model with non-constant zero-inflation.

The degree of zero-inflation is controlled by δ_0. All simulation experiments are run for two levels of zero-inflation, 10% and 50% respectively. These values were chosen to reflect the range of modest to substantial zero-inflation typically encountered in applications.

Count data models are unlikely to be of use if the proportion of excess zeros is higher. To obtain 10% zero-inflation, δ_0 is set equal to -2.197 in the constant, and equal to -2.564 in the non-constant zero-inflation set-up (where 10% is the average). A value of $\delta_0 = 0$ results in 50% zero-inflation in both cases.

The CEF of the Poisson part of the model, λ, contains a random component v, which is distributed independently of x as $Normal(\mu, \sigma^2)$. The true data generating process, unconditional on v, is therefore a zero-inflated Poisson-log-normal model. The random term v can be best thought of as an omitted variable that affects the mean of the count but is unobserved to the econometrician. Such unobserved heterogeneity, if unaccounted for or wrongly specified, leads to bias of zero-inflated models. To illustrate the amount of bias in finite samples, estimators are obtained from the zero-inflated Poisson and zero-inflated negative binomial models, in addition to PQL.

In the limit, $\sigma^2 = 0$ and there is no unobserved heterogeneity. In this case, the data generating process is indeed ZIP with $\lambda = \exp(0.5 + x)$ and zero-inflation of 10% or 50%. This experiment will allow us to compare the efficiency of PQL relative to the correctly specified and, thus, asymptotically efficient ZIP ML estimator. The scenario of no unobserved heterogeneity is quite unlikely in practice. Unobserved heterogeneity introduces overdispersion in the Poisson part of the model, since

$$\mathrm{Var}(y^* | \, x) = \mathrm{E}_v[\mathrm{Var}(y^*|x, v)] + \mathrm{Var}_v[\mathrm{E}(y^*|x, v)] = \mathrm{E}(y^* | \, x) + \mathrm{E}(y^* | \, x)^2 (e^{\sigma^2} - 1) e^{-\mu - \frac{1}{2}\sigma^2} \tag{3.9}$$

We parameterize μ and σ^2 in two different ways. If σ^2 is constant, it follows from (3.9) that the variance is a quadratic function of the mean. If, in addition, $\mu = -0.5\,\sigma^2$, then the CEF of the parent model is $\mathrm{E}(y^*|x) = \exp(\alpha + \beta x)$. Specifically, we assume $v \sim N(-0.5, 1)$. For this data generating process, we expect the ZINB to behave quite satisfactorily as the misspecification is limited to higher order moments, not mean and variance. The ZIP model by contrast assumes equality between mean and variance and is thus unlikely to produce

good results. The PQL estimator is robust to this kind of misspecification and should work well

A sparse way of obtaining different variance functions for y^* is by parametrizing μ and σ^2 as follows:

$$\sigma^2 = \ln\{1 + c\exp[(k-1)(\alpha + \beta x)]\} \qquad \mu = -0.5\,\sigma^2$$

The parameter k controls the nonlinearity of the variance function, while c is a free overdispersion parameter. In our third set-up, $c = 2$ and $k = -1$, implying a variance function with additive constant

$$\mathrm{Var}(y^*|x) = \mathrm{E}(y^*|x) + 2$$

The corresponding variance-to-mean ratio is now hyperbolic. In this case, all three estimators – ZIP, ZINB and PQL – only specify the first moment correctly. This should not matter for PQL but lead to bias for ZIP as well as ZINB.

For all setups two sample sizes with 100 and 1000 observations, respectively, were considered. The number of replications was 10,000 for every data generating process. The Monte Carlo study was programmed in STATA/MP 10.1; program code and full output are available on request.

3.4.2 Results

The results of the three simulation setups are displayed in Tables 3.1 to 3.3. Every table is divided into two panels, a left-hand side panel containing the results for the case of the zero-inflation parameter π being constant, and a right-hand side panel presenting results for a logit specification of π. We concentrate on the main parameter of interest, the semi-elasticity β whose true value is 1. The main entries in the tables are the mean of the QL and ML estimates $\hat{\beta}$ over the 10,000 replications. The numbers in parentheses give the standard deviations.

Table 3.1 shows the results for the first setup in which the data generating process is a ZIP. Not surprisingly, the ZIP estimates are very close to the true value on average, regardless of whether the sample size is 100 or 1000, and whether the degree of zero-inflation is 10% or 50%. However, the PQL performs well also. This is certainly true for the larger sample size and whenever the degree of zero-inflation is modest. The combination of small sample size and high degree of zero-inflation leads to some bias. For example, in the generating process with logit type zero-inflation, the PQL on average underestimates the true semi-elasticity by around 4 percent.

The more conspicuous difference between ZIP ML and PQL lies in their relative efficiency. The efficiency gains of using the correctly specified ZIP estimator are substantial. For instance, the standard deviation of the estimator is reduced by around one third to one half when passing from PQL to the ZIP estimate of β, depending on the severity of zero-inflation. The relative efficiency gains of ZIP in relation to PQL remain un changed as the sample size shrinks, and are in the same order of magnitude for constant and non-constant zero-inflation.

Figure 3.1 shows normal quantile-quantile (QQ) plots for the empirical distribution of $\hat{\beta}$ for $n = 100$. The values were centered at the mean value and normalized by the empirical standard deviation. The long-dashed and short-dashed lines correspond to the ZIP and PQL estimators, respectively. The solid 45^o line indicates where the points of a standard normal distribution in a normal QQ-plot would fall. The plots show that the normal approximation is quite satisfactory for all models even in small samples. Hence, approximate confidence intervals and t-statistics based on a limiting standard normal distribution should work well in practice.

Table 3.2 contains results obtained under the second setup where unobserved heterogeneity is causing the parent model to exhibit quadratic overdispersion. Irrespectively of the sample size, constant zero-inflation leads to large biases of the ZIP estimator. In contrast, the PQL estimates are practically unaffected by the presence of unobserved het-

erogeneity. The new DGP does make itself noticed in the larger standard deviations of the PQL estimates. As ZINB correctly specifies the CEF and the variance function, the table additionally includes results from this estimator. Not surprisingly, ZINB estimates β closely.

Results for non-constant zero-inflation in the right-hand side panel of Table 3.2 tell a similar story. As before, inconsistency of ZIP is reflected in substantial biases in all reported mean estimates, which are of the order of -15% to -20%. Estimation of the model with ZINB yields good results as would be expected. PQL estimates perform equally well in this setting. The efficiency advantage of ZINB over PQL is about one third in most cases.

In Table 3.3 the data are drawn from a process with additive overdispersion of y^*, so that both ZIP and ZINB only specify the CEF correctly. ZIP estimation again yields estimators that are not consistent for the true value of β in any of the entries of the table, with biases ranging from -5% to -12%. The ZINB estimator does not work well either, in particular for data with a high degree of zero-inflation, where the downward bias is up to 7%. By contrast, the performance of PQL is much better throughout. For the large sample, the discrepancy between the mean estimate and the true value of the parameter is always under 1%.

To summarize, the results from the Monte Carlo experiments in this section demonstrate the robustness of the PQL estimator of semi-elasticities in zero-inflated, finite samples, and the biases that can arise when using its two most common ZI competitors.

3.5 Application: demand for physician services

We illustrate PQL estimation of a count model with logit zero-inflation in an application related to health economics. In particular, the goal is to estimate how health insurance and other socio-demographic characteristics affect the frequency of doctor visits. The dataset is identical to the one used in Cameron and Trivedi (1986). The sample of 5190 individuals

is extracted from the Australian Health Survey 1977-78. The dependent variable is the number of consultations with a doctor or specialist in the two-week period prior to the interview. The mean is 0.302, the variance 0.637. Further details, and a motivation of the selection of explanatory variables, are given in Cameron and Trivedi (1986) and the references quoted therein.

Regressors include demographics (sex, age, age squared), income, various measures of health status (number of reduced activity days (actdays); general health questionnaire score (hscore); recent illness; two types of chronic conditions (chcond1, chcond2)), and three types of health insurance coverage (levyplus, freepoor, freerepat - the former representing a higher level of coverage and the latter two a basic level supplied free of charge).

Table 3.4 contains the regression results for the PQL estimator (in the first two columns) as well as for the fully parametric ZIP (in columns 3 and 4) and ZINB (in columns 5 and 6) models. In each case, all regressors enter both the logit model for zero-inflation and the log-linear CEF of the parent model. Their interpretation is accordingly one of changes in log-odds and semi-elasticities, respectively. A likelihood ratio test between ZIP and ZINB clearly favors the latter, an indication of the presence of unobserved heterogeneity and overdispersion. This does not mean, however, that the ZINB is the "right" model. If not, the estimator is inconsistent.

It is reassuring, therefore, that the parameter estimates are quite insensitive to the choice of specification in many instances, but there are exceptions. For instance, the ZINB model detects no statistically significant effect of having a chronic health condition in either part of the model. Under PQL, the second indicator has large negative and statistically significant effect on the probability of an extra zero and thus increases the expected number of visits. Inferences from PQL and ZINB also differ regarding insurance status. "freepoor" and "levyplus" are statistically significant in the ZINB but not so in the PQL model, suggesting some caution in interpreting these effects.

3.6 Concluding remarks

The main quantities of interest in most count data applications are the conditional expectation function and its semi-elasticities with respect to some regressors. For instance, all applications cited in the introduction without exception limited the discussion of their estimation results to these CEF effects. This paper proposed a new approach based on Poisson Quasi-Likelihood estimation as a way to estimate these quantities without having to specify more than the CEF, as opposed to the full distribution as is necessary with the traditional ZIP and ZINB ML estimators.

Zero-inflation can either be generated by a constant factor or else by a binary stochastic process. In the first case, simple estimation of the standard Poisson regression model yields consistent estimates of the semi-elasticities of the mean with respect to the independent variables. In the second case, a modification of the mean function is needed. In general, however, estimation of the parameters needed to estimate the conditional expectation and semi-elasticities is straightforward, as was illustrated in a set of Monte Carlo experiments.

The advantage of using PQL over ZIP and ZINB is its robustness to misspecification. Given the pervasive uncertainty about the data generating processes in practice, using estimators for ZI models seems unwise if concerns about bias from higher order misspecification exist. The relatively mild misspecifications of the DGP presented in the Monte Carlo experiments frequently resulted in noticeable biases, suggesting that PQL may be the better choice for estimating ZI models compared to ZI ML estimators in the absence of strong a priori information about the DGP. This conclusion will be more the compelling the larger the data set at hand.

References

Albert A. and J.A. Anderson (1984), On the Existence of Maximum Likelihood Estimates in Logistic Regression Models, *Biometrika*, **71**, 1-10.

Ateca-Amestoy, Victoria (2008), Determining heterogeneous behavior for theater attendance, *Journal of Cultural Economics*, **32**, 127-151.

Cameron, A. Colin, and Pravin K. Trivedi (1998), *Regression Analysis of Count Data*, Cambridge, MA: Cambridge University Press.

Campolieti, Michele (2002), The recurrence of occupational injuries: Estimates from a zero-inflated count model, *Applied Economics Letters*, **9**, 595-600.

Clark Ken, Simon A. Peters, and Mark Tomlinson (2005), The determinants of lateness: Evidence from British workers, *Scottish Journal of Political Economy*, **52**(2), 282-304.

Gourieroux, Christian, Alain Monfort and Alain Trognon (1984a), Pseudo Maximum Likelihood Methods: Theory, *Econometrica*, **52**, 681-700.

Gourieroux, Christian, Alain Monfort and Alain Trognon (1984b), Pseudo Maximum Likelihood Methods: Application to Poisson models, *Econometrica*, **52**, 701-721.

Hall, Daniel B., and Jing Shen, Robust estimation for zero-inflated Poisson regression, forthcoming in *Scandinavian Journal of Statistics*, published online September 27, 2009, DOI: 10.1111/j.1467-9469.2009.00657.x

Heitmueller, Axel (2004), Job mobility in Britain: Are the Scots different? Evidence from the BHBS, *Scottish Journal of Political Economy*, **51**(3), 329-358.

Ho, Woon-Yee, Peiming Wang and Joseph D. Alba (2009), Merger and acquisition FDI, relative wealth and relative access to bank credit: Evidence from a bivariate zero-inflated count model, *International Review of Economics and Finance*, **18**, 26-30.

Lambert, Diane (1992), Zero-inflated Poisson regression with an application to defects in manufacturing, *Technometrics*, **34**, 1-14.

List, John A. (2001), Determinants of securing academic interviews after tenure denial: evidence from a zero-inflated Poisson model, *Applied Economics*, **33**, 1423-1431.

Mullahy, John (1986), Specification and Testing of Some Modified Count Data Models, *Journal of Econometrics*, **33**, 341-365.

Papadopoulos, Georgios, and Joao M.C. Santos Silva (2008), Identification Issues in Models for Underreported Counts, University of Essex, Discussion Paper No. 657.

Sheu, Mei-Ling, Teh-Wei Hu, Theodore E. Keeler, Michael Ong and Hai-Yen Sung (2004), The effect of major cigarette price change on smoking behavior in California: a zero-inflated negative binomial model, *Health Economics*, **13**, 721-791.

Stephan, Paula E., Shiferaw Gurmu, Albert J. Sumell and Grant Black (2007), Who's patenting in the university? Evidence from the survey of doctorate recipients, *Economics of Innovation and New Technology*, **16**(2), 71-99.

White, Halbert (1982), Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, **50**, 1-25.

Winkelmann, Rainer, and Klaus F. Zimmermann (1993), Poisson Logistic Regression, University of Munich, Working Paper No. 93-18.

Winkelmann, Rainer (2008), *Econometric Analysis of Count Data*, fifth edition, Berlin: Springer.

Zahran, Sammy, Samuel D. Brody, Praveen Maghelal, Andrew Prelog and Michael Lacy (2008), Cycling and walking: Explaining the spatial distribution of healthy modes of transportation in the United States, *Transportation Research Part D*, **13**, 462-470.

Table 3.1: Estimated semi-elasticities – No overdispersion

		Constant zero-inflation		Logit zero-inflation	
Estimator		10%	50%	10%	50%
ZIP	n=100	1.0012	1.0032	0.9981	1.0024
		(0.0757)	(0.1155)	(0.0795)	(0.1325)
	n=1000	1.0000	1.0003	1.0003	1.0002
		(0.0216)	(0.0310)	(0.0221)	(0.0357)
PQL	n=100	0.9986	0.9822	0.9763	0.9628
		(0.0977)	(0.2207)	(0.1265)	(0.2469)
	n=1000	0.9999	0.9975	0.9925	0.9939
		(0.0310)	(0.0743)	(0.0449)	(0.0873)

Notes: Entries are the average estimates over 10,000 replications. Standard deviations in parenthesis. True value: $\beta = 1$.

Table 3.2: Estimated semi-elasticities – Quadratic overdispersion

		Constant zero-inflation		Logit zero-inflation	
Estimator		10%	50%	10%	50%
ZINB	n=100	1.0001	0.9993	0.9382	0.9580
		(0.1791)	(0.2692)	(0.1834)	(0.2779)
	n=1000	0.9993	1.0026	0.9515	0.9865
		(0.0567)	(0.0823)	(0.0604)	(0.0852)
ZIP	n=100	0.8578	0.8329	0.8212	0.8006
		(0.2464)	(0.3273)	(0.2436)	(0.3152)
	n=1000	0.8837	0.8653	0.8624	0.8520
		(0.1034)	(0.1326)	(0.0952)	(0.1263)
PQL	n=100	0.9743	0.9437	0.9378	0.9400
		(0.2344)	(0.3397)	(0.2597)	(0.3640)
	n=1000	0.9962	0.9898	0.9827	0.9837
		(0.0955)	(0.1347)	(0.0976)	(0.1448)

Notes: See Table 1

Table 3.3: Estimated semi-elasticities – Additive overdispersion

		Constant zero-inflation		Logit zero-inflation	
Estimator		10%	50%	10%	50%
ZINB	n=100	0.9803	0.9449	0.9592	0.9547
		(0.1431)	(0.2022)	(0.1579)	(0.2466)
	n=1000	0.9745	0.9259	0.9744	0.9316
		(0.0533)	(0.0676)	(0.0528)	(0.0816)
ZIP	n=100	0.9385	0.9060	0.9266	0.8801
		(0.1393)	(0.2007)	(0.1533)	(0.2504)
	n=1000	0.9321	0.9112	0.9321	0.8849
		(0.0450)	(0.0581)	(0.0441)	(0.0753)
PQL	n=100	1.0015	0.9866	0.9770	0.9783
		(0.1258)	(0.2438)	(0.1622)	(0.2803)
	n=1000	0.9922	0.9985	0.9921	0.9958
		(0.0539)	(0.0807)	(0.0537)	(0.0979)

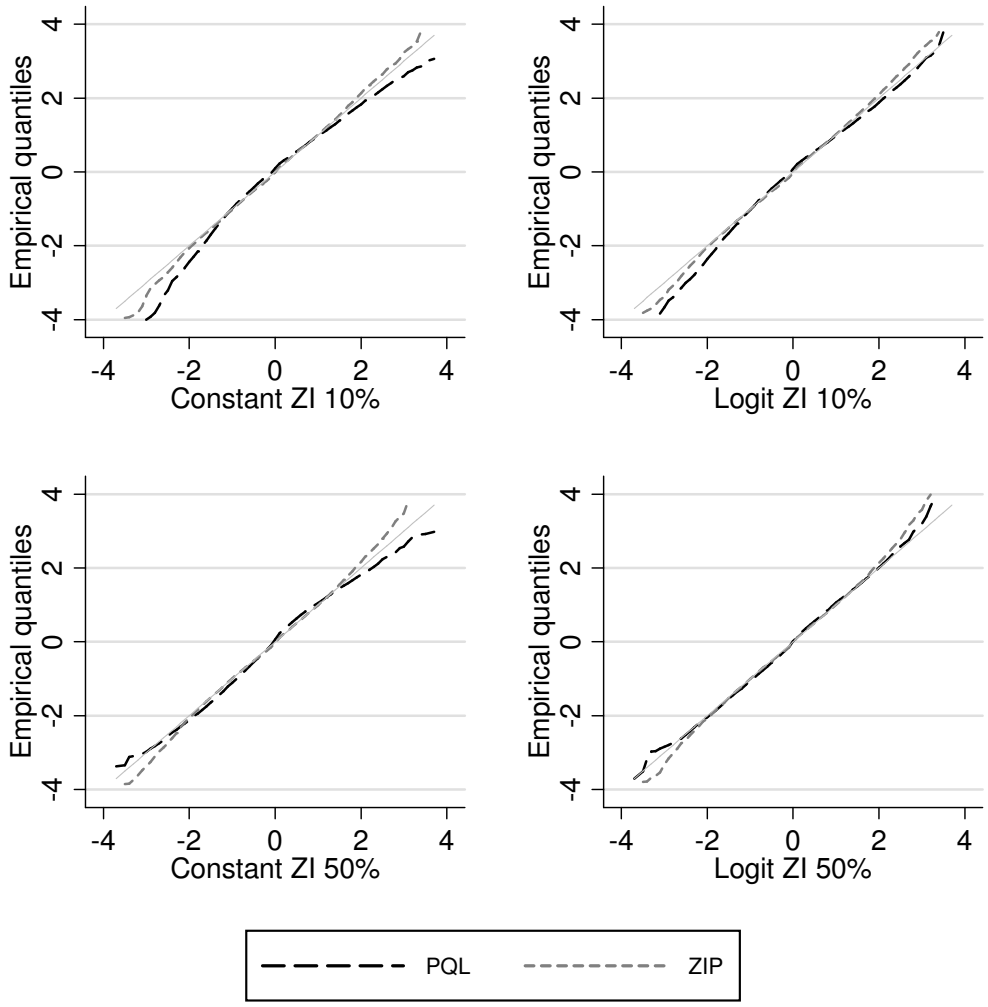
Notes: See Table 1.

Table 3.4: Zero-Inflation models for number of doctor consultations (n=5190)

	PQL		ZIP		ZINB	
Variable	ZI	Parent	ZI	Parent	ZI	Parent
Sex	-0.275	0.003	-0.488***	-0.027	-0.592***	0.010
	(0.228)	(0.135)	(0.171)	(0.072)	(0.228)	(0.084)
Age $\times 10^{-2}$	8.864**	3.784*	10.496***	3.128**	10.677**	2.103
	(3.986)	(2.212)	(3.271)	(1.297)	(4.386)	(1.541)
Age squared $\times 10^{-4}$	-10.611*	-3.882*	-13.337***	-3.409**	-13.821***	-2.187
	(4.379)	(2.341)	(3.690)	(1.374)	(5.002)	(1.639)
Income	-0.269	-0.288	-0.437*	-0.295***	-0.365	-0.214
	(0.349)	(0.203)	(0.264)	(0.113)	(0.346)	(0.133)
Levyplus	-0.381	-0.032	-0.433**	-0.034	-0.640**	-0.095
	(0.253)	(0.158)	(0.197)	(0.096)	(0.264)	(0.114)
Freepoor	0.278	-0.385	0.308	-0.377	0.111	-0.481*
	(0.830)	(0.512)	(0.508)	(0.239)	(0.659)	(0.283)
Freerepat	-0.974**	-0.254	-1.149***	-0.215*	-1.375***	-0.189
	(0.339)	(0.202)	(0.305)	(0.117)	(0.447)	(0.140)
Illness	-0.345**	0.002	-0.416***	0.049**	-0.672***	0.052*
	(0.092)	(0.045)	(0.081)	(0.025)	(0.156)	(0.029)
Actdays	-1.114**	0.047***	-1.256***	0.083***	-1.787***	0.104***
	(0.198)	(0.014)	(0.238)	(0.006)	(0.653)	(0.008)
Hscore	-0.080*	0.016	-0.097**	0.018	-0.105*	0.023*
	(0.043)	(0.020)	(0.039)	(0.011)	(0.056)	(0.014)
Chcond1	-0.242	-0.078	-0.127	-0.013	-0.119	-0.000
	(0.262)	(0.164)	(0.199)	(0.092)	(0.279)	(0.108)
Chcond2	-0.754**	-0.144	-0.604**	-0.034	-0.489	0.055
	(0.352)	(0.180)	(0.306)	(0.103)	(0.414)	(0.121)
Const.	1.452**	-0.618	0.786	-1.050***	0.622	-1.233***
	(0.739)	(0.472)	(0.572)	(0.255)	(0.753)	(0.296)
γ^{-1}						-0.578
						(0.080)
Log-likelihood			-3174.2		-3107.6	

Notes: Standard errors in parentheses. ***, **, * denote statistical significance at the 1%, 5%, 10% significance levels, respectively.

Figure 3.1: Normal QQ-plots for semi-elasticities



Notes: Standardized empirical quantiles plotted against theoretical normal quantiles. Data: 10'000 estimates of β from setup 1 (no overdispersion) with sample size = 100.

Chapter 4

A causal interpretation of extensive and intensive margin effects in generalized Tobit models

This chapter is available as Working Paper No. 1012, *SOI Working Papers Series*, Department of Economics, University of Zurich.

Acknowledgements: I am indebted to Rainer Winkelmann for extensive discussions and many helpful suggestions which improved this article. I would like to thank Colin Cameron for an insightful discussion, as well as Gregori Baetschmann and Stefan Boes for useful inputs. I am also grateful to conference participants at the European Trade Study Group Conference in Fribourg, Switzerland, and the Midwest Econometrics Group Annual Meeting in St. Louis, MO, USA, as well as seminar participants at the University of Zurich, for various comments. Opinions and errors are mine.

4.1 Introduction

Many outcomes of interest in economics are nonnegative and have a cluster of observations at the value zero. Prominent examples include working hours, health care demand, and expenditure data. More generally, variables with these features are referred to as corner solution outcomes (Wooldridge, 2002), which suggests the idea of utility maximization under constraints where both interior and corner solutions occur, for instance due to kinks in budget constraints.

Researchers analyzing effects of variables on corner solution outcomes frequently take interest in decomposing the effect into the part attributable to individuals starting to participate (called *extensive margin*), and the part attributable to already participating individuals (called *intensive margin*). The decomposition used is algebraically straightforward as it is based on factoring the expectation of the corner solution variable, say $\mathrm{E}(Y)$, into the participation probability $\Pr(Y > 0)$ and the conditional expectation $\mathrm{E}(Y|Y > 0)$ (McDonald and Moffitt, 1980). The extensive margin is driven by the participation effect [PE], the change in the probability to participate; the intensive margin is driven by the conditional-on-positives effect [COP], the change in the outcome given participation.

In contrast to the simplicity of the mechanical aspect, endowing the decomposition with a causal interpretation is substantially more problematic. For instance, recent work framing the problem in terms of Rubin's potential outcomes model has pointed out that COP effects do not measure the impact of a treatment on participating individuals; rather, they are hopelessly contaminated by a sort of selection bias, even in experimental settings (Angrist, 2001; Angrist and Pischke, 2009). An apparent solution is resorting to the interpretation of effects on underlying, latent variables such as in censored regression and sample-selection models, where causal interpretation is feasible. However, as these authors and others emphasize (cf. Dow and Norton, 2003), latent outcomes are artificial and lack a meaningful interpretation in corner solution contexts.

In this article, I propose a conceptually different decomposition of the effect into ex-

tensive and intensive margins. It is based on the joint distribution of potential outcomes, which ensures that the resulting parts are meaningful in a causal sense. Indeed, the new decomposition succeeds in representing the total effect as an average of the treatment effects for interesting subgroups of the population: those induced to participate by the treatment, and those participating regardless of it. Like the conventional decomposition, this one is not identified nonparametrically, although sharp bounds can be derived for the average treatment effect of the population subgroups. Imposing some more structure point-identifies the decomposition. Examples include a class of generalized Tobit models, which are widely used in applied research. For these models, the differences between decompositions can be major.

An application to the gravity model of trade compares the two decompositions in a real-world setting. The decomposition of trade effects into extensive and intensive margins is an issue of ongoing interest in the recent empirical trade literature (Felbermayr and Kohler, 2006; Helpman, Melitz and Rubinstein, 2008; Liu, 2009). Here, I estimate the effect of a hypothetical reduction in entry regulation costs on bilateral trade flows, and decompose it into country margins. The estimates suggest that the usual decomposition overstates the contribution of the extensive margin by around 15%.

Practitioners confronted with limited dependent variables in many diverse fields of applied economics and other social sciences will find this article to be of interest. Examples of work featuring the decomposition of effects into extensive and intensive margins include estimates of the effect of benefit-receiving on food expenditure (Hastings and Washington, 2010), the effect of various variables on intra- and inter-firm trade (Co, 2010), the effect of employer contributions on employee pension savings (Engelhardt and Kumar, 2007) the effect of worker productivity or unionization on working overtime (Sousa-Poza and Ziegler, 2003, and Trejo, 1993, respectively), the effect of managers' tax evasion preferences on underreporting corporate income (Joulfaian, 2000), the effect of various regressors on youth unemployment (Caspi et al. 1998), and the effect of health knowledge on health outcomes

(Kenkel, 1991). The list is neither complete nor representative, but it is suggestive of the widespread use of the decomposition in corner solution applications.

This article contributes to the growing recent literature on treatment effects for limited dependent variables (Aakvik, Heckman and Vytlacil, 2005; Chen, 2010; Chiburis, 2010; Fan and Wu, 2010). Since its emphasis lies on conceptual definition of objects of interest and interpretation, it is close in spirit to Angrist (2001). The representation of treatment effects as weighted sums of population groups is influenced by Angrist and Imbens (1994). As this framework is expressible in latent index models (Vytlacil, 2002), latent index representations with binary endogenous variables obtain expressions that resemble the ones presented below. Conceptually, they are quite different, because in the endogenous treatment literature, the population subgroups are defined by their potential treatment status in response to an instrument, while here the groups are defined by their potential outcome status in response to the (exogenous) treatment.

The plan for the article is this: the next section reprises the Angrist-Pischke "bad COP"-critique, presents the alternative decomposition and discusses its nonparametric identification. Section 4.3 exemplifies the new approach for some common Tobit-type models, and section 4.4 provides the application to the gravity model of trade. Section 4.5 contains a concluding discussion.

4.2 Corner solutions and potential outcomes

Consider the causal effect of a binary treatment T_i on the corner solution variable $Y_i \geq 0$ for individuals $i = 1, \ldots, N$. Let Y_{1i} denote the outcome for i if i received the treatment, i.e. $T_i = 1$, and Y_{0i} if $T_i = 0$, so that as usual

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})T_i$$

The focus here will be on the causal treatment effect $Y_{1i} - Y_{0i}$. Assume that the data comes from an ideal randomized controlled trial, so that assignment to treatment is random and

compliance is perfect. Then, T_i is independent of (Y_{1i}, Y_{0i}), and the average treatment effect [ATE] $\mathrm{E}(Y_{1i}-Y_{0i})$ can be obtained from the prima-facie contrast $\mathrm{E}(Y_i|T=1)-E(Y_i|T=0)$. Using

$$\mathrm{E}(Y_i|T_i) = \Pr(Y_i > 0|T_i)\mathrm{E}(Y_i|Y_i > 0, T_i)$$

this contrast can be written as

$$\begin{aligned}
&\mathrm{E}(Y_i|T=1) - E(Y_i|T_i=0) = \\
&\quad \big\{ \Pr(Y_i > 0|T_i=1) - \Pr(Y_i > 0|T_i=0)\big\}\mathrm{E}(Y_i|Y_i > 0, T_i=1) \\
&\quad +\big\{\mathrm{E}(Y_i|Y_i > 0, T_i=1) - \mathrm{E}(Y_i|Y_i > 0, T_i=0)\big\} \Pr(Y_i > 0|T_i=0) \qquad (4.1)
\end{aligned}$$

This is the usual decomposition applied to limited dependent variables like Y_i in Tobit (Tobin, 1958) or Cragg (1971) models (McDonald and Moffitt, 1980; cf. also the standard graduate textbooks by Cameron and Trivedi, 2005, Greene, 2008, and Wooldridge, 2002). The first term after the equality sign is the extensive margin effect, which weights the PE —the term in curly brackets— by the expected Y_i conditional on participation; the second term is the intensive margin effect, which weights the COP (in curly brackets) by the probability of participation given $T_i = 0$. Angrist (2001) and Angrist and Pischke (2009) suggest rewriting COP in terms of potential outcomes as

$$\begin{aligned}
&\mathrm{E}(Y_i|Y_i > 0, T_i=1) - \mathrm{E}(Y_i|Y_i > 0, T_i=0) \\
&\quad = \mathrm{E}(Y_{1i}|Y_{1i} > 0, T_i=1) - \mathrm{E}(Y_{0i}|Y_{0i} > 0, T_i=0) \\
&\quad = \mathrm{E}(Y_{1i}|Y_{1i} > 0) - \mathrm{E}(Y_{0i}|Y_{0i} > 0) \\
&\quad = \mathrm{E}(Y_{1i} - Y_{0i}|Y_{1i} > 0) + \big\{\mathrm{E}(Y_{0i}|Y_{1i} > 0) - \mathrm{E}(Y_{0i}|Y_{0i} > 0)\big\} \qquad (4.2)
\end{aligned}$$

The second equality follows by independence of T_i from Y_{1i} and Y_{0i}. If only independence from Y_{0i} was assumed, as in Angrist's (2001) formulation, (4.2) would need to be written conditional on $T_i = 1$. This has no bearing on the present argument. As can be seen from the terms after the third equality, COP is composed of two terms. The first, $\mathrm{E}(Y_{1i}-Y_{0i}|Y_{1i} >$

0), can be given a causal interpretation. It is the treatment effect for the subpopulation of individuals having positive Y_i when $T_i = 1$. The second term, $\mathrm{E}(Y_{0i}|Y_{1i} > 0) - \mathrm{E}(Y_{0i}|Y_{0i} > 0)$, can be understood as a form of selection bias. The selection bias in COP arises because treatment has an effect on the composition of the group with $Y_i > 0$: The interest lies in those with $Y_{1i} > 0$, but the COP contrast also involves the group $Y_{0i} > 0$ which might be a super- or sub-set of the group $Y_{1i} > 0$, but not the same unless treatment has no effect on the participation probability. Thus, the analysis in (4.2) implies that using a decomposition like (4.1) cannot identify a causal effect even in ideal settings like a randomized controlled trial.

However, the more fundamental problem is that the first term in (4.2) is not an object of direct interest in a decomposition into extensive and intensive margins. The ATE for individuals with $Y_{1i} > 0$ mixes the ATE for the two population groups the decomposition set out to discriminate, the ones participating even without treatment and the ones participating because of the treatment.

4.2.1 Decomposition based on joint outcomes

Consider the following classification of individuals into non-overlapping and exhausting groups based on their *joint* distribution of potential outcomes, (Y_{0i}, Y_{1i}):

Group	Name	Potential outcomes
NP	Non-participants	$(Y_{0i} = 0, Y_{1i} = 0)$
S_1	Switchers	$(Y_{0i} = 0, Y_{1i} > 0)$
S_2	Switchers	$(Y_{0i} > 0, Y_{1i} = 0)$
P	Participants	$(Y_{0i} > 0, Y_{1i} > 0)$

Basing the definition of intensive and extensive margin effects on these groups clarifies their meaning substantially. The intensive margin effect is the contribution to the ATE of group P. Similarly, the extensive margin is the ATE contribution of switchers, i.e. those

changing their participation status (groups S_1 and S_2). These are the objects of interest when decomposing causal effects into extensive and intensive margins; when researchers write about them, it is this what they mean (although they rarely state it so explicitly). For instance, take the labor economics example of working hours. The effect of a policy intervention increasing average working hours in the economy can be decomposed into

– the average change in hours worked of those working regardless of the intervention,

– plus the average hours worked by those joining the workforce because of the intervention,

– minus the average hours worked by those leaving the workforce because of the intervention,

the groups being weighted by their population fraction.

Often researchers choose models which possess some monotonicity assumption on the way treatment affects outcomes (Manski, 1997). This can lead to the elimination of one group out of S_1 and S_2. For instance, a strong monotone treatment response assumption states that the causal effect $Y_{1i} - Y_{0i}$ is either nonnegative or nonpositive for all i. In the working hours example, this means that if the policy increased working hours of workers, no one is induced to leave the workforce (group S_2 is ruled out). Such an assumption is embedded in the Tobit model. The monotone treatment response assumption can be weaker and still eliminate one group. Tautologically, it is sufficient that the causal effect is either positive for all i with $Y_{0i} = 0$ or $Y_{1i} = 0$, or negative. This assumption is implicit in Cragg's (1971) model. Often such assumptions are motivated by economic theory, and for many applications it might be plausible to impose them. Finally, the effect for individuals in group NP is always zero.

Thus, formally, the decomposition of the ATE based on the joint distribution of potential outcomes is

$$\begin{aligned} \mathrm{E}(Y_i|T=1) - \mathrm{E}(Y_i|T=0) &= \mathrm{E}_{Y_{1i},Y_{0i}}\left[\mathrm{E}(Y_{1i}-Y_{0i})|Y_{1i},Y_{0i}\right] \qquad (4.3) \\ &= \mathrm{E}(Y_{1i}|Y_{0i}=0,Y_{1i}>0)\Pr(Y_{0i}=0,Y_{1i}>0) \\ &+ \mathrm{E}(-Y_{0i}|Y_{0i}>0,Y_{1i}=0)\Pr(Y_{0i}>0,Y_{1i}=0) \\ &+ \mathrm{E}(Y_{1i}-Y_{0i}|Y_{0i}>0,Y_{1i}>0)\Pr(Y_{0i}>0,Y_{1i}>0) \end{aligned}$$

As before, the left-hand side of (4.3) corresponds to $\mathrm{E}(Y_{1i}-Y_{0i})$ because of randomized treatment assignment. The expectation over (Y_{1i},Y_{0i}) is with respect to the four events NP, S_1, S_2 and P. The last term is the intensive margin effect [IME], the first two are the extensive margin effect [EME], though as noted most models used in the literature will eliminate one of these.

4.2.2 Nonparametric identification

A comparison between (4.1) and (4.3) shows that they are distinct decompositions even under the monotone treatment response assumption. To highlight the difference, section 4.3 applies (4.1) and (4.3) to some of the most common corner solution response models in the literature. In those models, the decomposition based on joint potential outcomes is identified because considerable structure is imposed through functional form restrictions. Alternative identifying assumptions are discussed in the concluding section. In this section, it is shown that the decomposition is not identified nonparametrically: Experimental data combined with a monotone treatment response assumption alone do not point-identify all the objects of interest involved in the decomposition. The considerations below use the weaker monotone treatment response assumption that the treatment effect is either positive for all switchers, or negative.

A more compact notation will facilitate the exposition. Define the population fractions of switchers $\pi^S \equiv \Pr(Y_{i0}=0,Y_{i1}>0)$ and participants $\pi^P \equiv \Pr(Y_{i0}>0,Y_{i1}>0)$. Simi-

larly, define mean potential outcomes of switchers $(\bar{Y}_0^S, \bar{Y}_1^S)$ and of participants $(\bar{Y}_0^P, \bar{Y}_1^P)$ (for instance, $\bar{Y}_0^S = \mathrm{E}(Y_{i1}|Y_{i0} = 0, Y_{i1} > 0)$). Then, the decomposition of the average treatment effect might be written as

$$\mathrm{ATE} = \pi^S \bar{Y}_1^S + \pi^P(\bar{Y}_1^P - \bar{Y}_0^P) = \pi^S \mathrm{ATE}^S + \pi^P \mathrm{ATE}^P \tag{4.4}$$

for the case $\bar{Y}_0^S = 0$, i.e. the case of group S_2 having mass zero. The reverse case $(\bar{Y}_0^S > 0, \bar{Y}_1^S = 0)$, i.e. group S_1 having mass zero, will not be considered — being symmetric, it gives no additional insights. Given the monotone treatment response assumption that one of the two cases holds, population regression of $D_i \equiv \mathbb{1}(Y_i > 0)$ on T_i reveals which it is: The difference $\mathrm{E}(D_i|T_i = 1) - \mathrm{E}(D_i|T_i = 0)$ corresponds to the difference between the S_1-fraction and the S_2-fraction in the population. Thus, if $\mathrm{E}(D_i|T_i = 1) > \mathrm{E}(D_i|T_i = 0)$, it must be that the S_2-fraction has mass zero.

Population regression of D_i on T_i can then be used to determine the population fractions of switchers and participants

$$\pi^S = \mathrm{E}(D_i|T_i = 1) - \mathrm{E}(D_i|T_i = 0) \quad \text{and} \quad \pi^P = \mathrm{E}(D_i|T_i = 1)$$

The term $\bar{Y}_0^P$ is also identified by the data; $\bar{Y}_0^P = \mathrm{E}(Y_i|D_i = 1, T_i = 0)$. The problem is identification of $\bar{Y}_1^S$ and $\bar{Y}_1^P$ for which only one quantity exists in the data, $\mathrm{E}(Y_i|D_i = 1, T_i = 1)$:

$$\mathrm{E}(Y_i|D_i = 1, T_i = 1) = \omega^S \bar{Y}_1^S + (1 - \omega^S)\bar{Y}_1^P, \quad \omega^S = \frac{\pi^S}{\pi^S + \pi^P}$$

Thus, it is impossible to disentangle them without making more assumptions; it follows that the decomposition is not identified nonparametrically — it is not possible to attribute a fraction of ATE to the extensive or intensive margin. However, since ATE, π^S and π^P are all identified, it is possible to derive sharp bounds for ATE^S and ATE^P using (4.4).

The bounds depend on the sign of ATE. Assume ATE < 0 first. Since $\mathrm{ATE}^S > 0$ (because S_2 is ruled out, as before), this means ATE^P must be negative. The domain of ATE^S is the positive real line $(0; \infty)$; the domain of ATE^P is the interval $(-\bar{Y}_0^P; 0)$.

Substituting the limits of these intervals into (4.4) the identification regions for the objects of interest reduce to

$$\text{ATE}^S \in \left(0, (\text{ATE} + \pi^P \bar{Y}_0^P)/\pi^S\right), \quad \text{ATE}^P \in \left(0, \text{ATE}/\pi^P\right)$$

The bounded regions are strictly smaller than the supports, and therefore informative.

Consider now ATE > 0. The data do not reveal the sign of the average treatment effect for participants. A strong monotone treatment response assumption would restrict it to be positive, so that ATE^P's domain would be $(0, \infty)$. In that case, a similar argument to the one above gives the following intervals for the ATE in the two population groups:

$$\text{ATE}^S \in \left(0, \text{ATE}/\pi^S\right), \quad \text{ATE}^P \in \left(0, \text{ATE}/\pi^P\right)$$

If one is unwilling to make this strong monotonicity assumption, the possibility of a negative ATE for participants has to be taken into account, which widens the domain of ATE^P; identification intervals are also widened in consequence but remain informative. If ATE^P < 0 this bounding strategy does not reduce ATE^P's domain which remains $(-\bar{Y}_0^P, 0)$. Combined with the previous result it follows that ATE^P $\in (-\bar{Y}_0^P, \text{ATE}/\pi^P)$, while ATE^S $\in (0, (\text{ATE} + \pi^P \bar{Y}_0^P)/\pi^S)$.

As in other partial identification settings (Manski, 2003), the availability of an additional discrete exogenous variable, say X_i, could tighten the upper bound for ATE^S further if X_i was related to the probability of being a switcher. Bounds for the average treatment effect could then be calculated for every value of the exogenous variable. The new upper bound for ATE^S resulting from adding the conditional-on-X_i bounds weighted by the mass points of X_i could be smaller than the ones given above. Similar arguments can be made for the bounds of participants.

4.3 Decomposing ATE in some common structural models

This section illustrates the decomposition based on joint potential outcomes for a class of models in which the objects of interest are point-identified.

4.3.1 Tobit model

The Tobit model (Tobin, 1958) is arguably the most popular model for corner solution dependent variables. It consists of three parts: a latent variable with a linear index structure, a distributional assumption for the error and an observation mechanism. These are

$$Y_i^* = \beta_0 + \beta_1 T_i + U_i \ , \quad U_i|, T_i \sim N(0, \sigma^2) \ , \quad Y_i = \max(0, Y_i^*) \tag{4.5}$$

Consider the case $\beta_1 > 0$, which imposes that Y_i is non-decreasing in T_i. The ATE in this model is

$$\mathrm{E}(Y_i|T_i = 1) - \mathrm{E}(Y_i|T_i = 0) = \Phi_1 \left(\beta_0 + \beta_1 + \sigma\phi_1/\Phi_1\right) - \Phi_0 \left(\beta_0 + \sigma\phi_0/\Phi_0\right)$$

where Φ_1, Φ_0 are the cdf of the standard normal distribution evaluated at $(\beta_0 + \beta_1)/\sigma$ and β_0/σ, respectively, and ϕ_1, ϕ_0 are the corresponding pdf. The conventional decomposition (4.1) would split this between extensive and intensive margin as follows

$$\widetilde{\mathrm{IME}} = (\beta_1 + \sigma\phi_1/\Phi_1 - \sigma\phi_0/\Phi_0)\Phi_0 \quad \widetilde{\mathrm{EME}} = (\Phi_1 - \Phi_0)(\beta_0 + \beta_1 + \sigma\phi_1/\Phi_1)$$

As was discussed before, it is difficult to assign a causal interpretation to $\widetilde{\mathrm{IME}}$ and $\widetilde{\mathrm{EME}}$. In the Tobit model, the distribution of potential outcomes of an individual is completely determined by her realization of the stochastic part U_i. Assume $\beta_1 > 0$, for concreteness. Then, individuals with U_i smaller than $-\beta_0 - \beta_1$ never have a positive Y_i; they constitute group NP (see Fig. 1). Similarly, if U_i lies between $(-\beta_0 - \beta_1)$ and $(-\beta_1)$, individuals are group S_1 switchers. (S_2 switchers, i.e. individuals dropping out of participation because

of treatment, are incompatible with the structure of the model when $\beta_1 > 0$.) $\widetilde{\text{EME}}$ correctly identifies the fraction of switchers ($\Phi_1 - \Phi_0$) but fails to attribute the correct ATE. Rather, it assigns to them the average Y_i in the population of switchers and participants. This overestimates their contribution, as switchers' U_i are in the bottom tail of the error distribution among those with $Y_{1i} > 0$. Their true ATE is $\beta_0+\beta_1+\text{E}(U_i|-\beta_0-\beta_1 \leq U_i < \beta_1)$. Thus the correction term is the expectation of a doubly-truncated normal variable. Table 4.1 contains the features of switchers and participants in the Tobit model. Multiplying the second by the fourth row gives the causal IME and EME.

Consider a numerical example to illustrate the difference which using the decomposition based on joint potential outcomes can make. Suppose the DGP is (4.5) with $\beta_0 = 0, \beta_1 = 1, \sigma^2 = 1$. Then the ATE is about 0.68. The conventional decomposition assigns about 0.24 to the intensive and 0.44 to extensive margin effect. In contrast, the decomposition into causally meaningful margins reveals that of the total ATE of 0.68, 0.5 is due to the intensive and only 0.18 due to the extensive margin effect. The intensive margin contribution, which was only 36% using the old decomposition, is thus really 73%.

Similarly stark discrepancies are possible in practice. For instance, McDonald and Moffitt's (1980) application examined the effect of a negative income tax on working hours reductions by estimating a Tobit model. Using their decomposition, it assigned 22% of the estimated reduction in working hours to the extensive margin. A follow-up article by Moffitt (1982) reevaluated the same data. In this article, he modified the Tobit model to account for a model of labor market frictions. Incidentally, this leads to the same formulas for the decomposition as the ones using the decomposition based on joint outcomes presented in Table 4.1. Applying this decomposition, he now concluded that the extensive margin was responsible not for 22%, but for 84% of the reduction. The present article shows that even in the absence or misspecification of the specific labor market frictions model postulated in Moffitt (1982), the causal extensive margin contribution is 84%.

Coming back to the numerical illustration from before, the example DGP can also be

used to illustrate the bounds discussed in the previous section. Here, the average treatment effect for participants, ATE^P, is 1 ($= \beta_1$), and the ATE for switchers, ATE^S, is about 0.54. A researcher ignoring the DGP and reluctant to make any assumptions on it can conclude that $\text{ATE}^P \in (0; 1.36)$ and $\text{ATE}^S \in (0; 2)$.

4.3.2 Selection and two-part models

There are several generalizations and alternatives to the Tobit model that are formulated in the same framework (Cragg, 1971; Heckman, 1979; Duan et al., 1983; Amemiya, 1985). One variant postulates

$$\begin{aligned} Y_i &= D_i \exp(\beta_0 + \beta_1 T_i + U_i) \\ D_i &= \mathbb{1}(\alpha_0 + \alpha_1 T_i + V_i) \end{aligned} \tag{4.6}$$

with $(U_i, V_i) \sim BVN(0, 0, \sigma^2, 1, \rho)$. The exponential transformation of the right-hand side of Y_i ensures positivity of Y_i. It is common to refer to model (4.6) as 'selection model' when the errors are correlated, and as 'two-part' model when errors are independent (Hay and Olsen, 1984). It is clear from (4.6) that here, in contrast to the Tobit model, the signs of extensive and intensive margin need not be the same, as they are driven by the signs of α_1 and β_1, respectively. Moreover, (4.6) models the participation decision (the equation for D_i) and the outcome conditional on participation as (potentially) driven by two separate errors (U_i and V_i). Analogously to the Tobit model, the population fraction of groups are $(\Phi_1 - \Phi_0)$ for switchers and Φ_0 for participants, where now $\Phi_1 = \Phi(\alpha_0 + \alpha_1)$ and $\Phi_0 = \Phi(\alpha_0)$.

Consider the two-part model first, i.e. assume correlation $\rho = 0$. The essential feature of the two-part model is that because the errors are independent, the conditional error expectation $\text{E}(\exp(U_i)|V_i) = \exp(0.5\sigma^2)$ is the same for switchers and participants. This means that both decompositions coincide: A switcher has an expected treatment effect of $\exp(\beta_0 + \beta_1 + 0.5\sigma^2)$ which is just $\text{E}(Y_i|Y_i > 0, T_i = 1)$, and participants experience a percental change of $\exp(\beta_1) - 1$. Or, in terms of the analysis in (4.2), the selection

bias in the COP effect vanishes in this model because $\mathrm{E}(Y_{0i}|Y_{1i} > 0) = \mathrm{E}(Y_{0i}|Y_{0i} > 0) = \exp(\beta_0 + 0.5\sigma^2)$.

The assumption of zero selection bias is unwarranted in most applications. While randomization prevents dependence between treatment and the errors, there is no experiment which could possibly break the potential dependence between U_i and V_i — and applications where the researcher can be certain that this dependence is absent seem difficult to envision.

Thus, consider the selection model which allows correlation between the errors. Since the model estimated in the application in the following section is a selection model with covariates, model (4.6) is rewritten to accommodate this feature. With covariates, model (4.6) is

$$Y_i = D_i \times \exp(X_i\beta + \beta_T T_i + U_i) \tag{4.7}$$

$$D_i = \mathbb{1}(Z_i\alpha + \alpha_T T_i + V_i \geq 0) \tag{4.8}$$

with $(U_i, V_i) \mid T_i, X_i, Z_i \sim BVN(0, 0, \sigma^2, 1, \rho)$. This distributional assumption implies that treatment and regressors are independent of the errors. Regressors are collected in two vectors Z_i and X_i with corresponding coefficient vectors α and β. No exclusion restriction is placed on covariates. In principle, they can be identical, disjoint or overlapping, although economic considerations will commonly lead to a set of overlapping, if not identical, variables.

The normality assumption implies a probit model for the decision to participate, $\Pr(D_i = 1|Z_i) = \Phi(Z_i\alpha + \alpha_T T_i)$. Then, for given Z_i, switchers are defined by values of V_i lying in the interval $S \equiv [-Z_i\alpha - \alpha_T; -Z_i\alpha)$, and participants by $V_i \in P \equiv [-Z_i\alpha; \infty)$. For observations with characteristics Z_i, the fractions of participants and that of switchers are

$$\Pr(V_i \in P) = \Phi(Z_i\alpha), \quad \Pr(V_i \in S) = \Phi(Z_i\alpha + \alpha_T) - \Phi(Z_i\alpha)$$

As seen previously, both decompositions assign the same value to the population fraction, but they differ in assigning average treatment effects for switchers and participants. For

switchers, ATE^S conditional on covariates is

$$\text{ATE}^S = \text{ATE}(X_i, Z_i, V_i \in S) = \\ \exp(X_i\beta + \beta_T + 0.5\sigma^2)\frac{\Phi(\sigma\rho + Z_i\alpha + \alpha_T) - \Phi(\sigma\rho + Z_i\alpha)}{\Phi(Z_i\alpha + \alpha_T) - \Phi(Z_i\alpha)} \tag{4.9}$$

the correction term is the doubly-truncated expectation $\text{E}(\exp(U_i)|V_i \in S)$. Instead, the standard decomposition uses the expectation of Y_i given $D_i = 1$ and $T_i = 1$ for ATE^S. The expectation of Y_i conditional on participation is (cf. Terza, 1998)

$$\text{E}(Y_i|D_i = 1, Z_i, X_i) = \exp(X_i\beta + \beta_T T_i + 0.5\sigma^2)\frac{\Phi(\sigma\rho + Z_i\alpha + \alpha_T T_i)}{\Phi(Z_i\alpha + \alpha_T T_i)} \tag{4.10}$$

where the correction term is the simple truncated expectation $\text{E}(\exp(U_i)|X_i, V_i > -Z_i\alpha - \alpha_T T_i)$. Essentially, this produces the same pattern of discrepancies between decompositions as in the Tobit model, although $|\rho| < 1$ will lessen the magnitude of the difference (with the two-part model being the limit case). In this model, the relative size of the conventional extensive margin effect ($\widetilde{\text{EME}}$) relative to the causal one (EME) depends solely on the linear index of the participation equation $Z_i\alpha$, α_T and $\sigma\rho$, but not on β and β_T:

$$\frac{\widetilde{\text{EME}}}{\text{EME}} = \left(1 - \frac{\Phi(Z_i\alpha)}{\Phi(Z_i\alpha + \alpha_T)}\right) \Bigg/ \left(1 - \frac{\Phi(\sigma\rho + Z_i\alpha)}{\Phi(\sigma\rho + Z_i\alpha + \alpha_T)}\right) \tag{4.11}$$

Thus, $\rho > 0$ ($\rho < 0$) implies that the conventional distribution overestimates (underestimates) the causal EM. Also, for a given value of ρ, the extent of the discrepancy is increasing in both $Z_i\alpha$ and α_T.

For participants, the conditional average treatment effect is

$$\text{ATE}^P = \text{ATE}(X_i, Z_i, V_i \in P) = (\exp(\beta_T) - 1)\exp(X_i\beta)\frac{\Phi(\sigma\rho + Z_i\alpha)}{\Phi(Z_i\alpha)} \tag{4.12}$$

while the conventional decomposition prescribes

$$\exp(X_i\beta + \beta_T + 0.5\sigma^2)\frac{\Phi(\sigma\rho + Z_i\alpha + \alpha_T)}{\Phi(Z_i\alpha + \alpha_T)}\Phi(Z_i\alpha) - \exp(X_i\beta + 0.5\sigma^2)\Phi(\sigma\rho + Z_i\alpha)$$

Unconditional ATE for switchers and participants can be obtained by taking expectations over the distribution of (Z_i, X_i), e.g. $\text{ATE}^S = \text{E}_{Z_i,X_i}[\text{ATE}(X_i, Z_i, V_i \in S)]$.

4.4 An application: The trade effect of reducing the number of bureaucratic firm-entry-regulation procedures

This section applies the new decomposition to an empirical trade model. Traditionally, the determinants of trade volumes were estimated in a single-equation, constant-elasticity gravity model (Santos Silva and Tenreyro, 2006; Feenstra, 2008). However, the large fraction of zeros in aggregated trade datasets spanning many countries has motivated a new strand of empirical literature which favors a two-equations model (Helpman, Melitz and Rubinstein, 2008). The first equation addresses the zeros directly by modeling trade participation. The second equation models trade flows conditional on participation. The trade volume equation is specified as a traditional gravity model. With these equations, explanatory variables can influence trade flows at two country margins, the extensive margin –the decision to trade– and the intensive margin –average trade flows of trading country-pairs. In this application, I will analyze the trade effect of a hypothetical policy intervention which would reduce the number of bureaucratic procedures needed to set up a business legally.

The empirical model is the generalized Tobit model (4.7)-(4.8). The indicator variable D_i declares the presence or absence of trade between a directed country-pair i, and the variable Y_i will denote its trade volume ($Y_i \geq 0$). The term "directed country-pair" means here that for every pair of countries there are two observations: the exports of the first to the second and vice versa. The vector of variables explaining the decision to trade are Z_i, the variables explaining the trade volume X_i, and unobserved variables (as well as pure randomness) in the participation and volume equations are V_i and U_i, respectively. The set of variables X_i and Z_i can contain distinct elements, in principle. Indeed, much of the theoretic motivation for the two-equations model comes from the idea that zero trade flows are due to the impossibility of overcoming fixed costs which are necessary to establish

trade (Hallak, 2006). This suggests that Z_i contains "fixed costs" and X_i "variable costs" of trading. In practice, however, the case seems less clear-cut as at the aggregate country-level the variables observed are not the "costs" directly, but rather rough proxies for them, such as distance between capital cities, which makes it hard to distinguish between fixed and variable costs. For instance, firm entry regulation measures such as the number of procedures should primarily be a fixed cost and not affect variable trading costs (Helpman, Melitz and Rubinstein, 2008). But Djankov et al. (2002) relate such costs to corruption and shadow economies which are likely to affect variable trade costs as well. Baranga (2009, fn. 9) provides an alternative argument against excluding firm entry regulation variables from X_i: "[A] country with higher regulatory barriers may also be more likely to be a higher tax environment, which would be expected to reduce the profitability of exporting at the intensive margin too. Countries with more regulation might also be more likely to use quantitative trade restrictions such as import or export licenses, or other non-tariff barriers, which would also affect the intensive margin, but are typically not controlled for." Thus, no exclusion restrictions will be placed on the variables here, so that $X_i = Z_i$. Finally, adopting the assumption of bivariate normal errors facilitates comparison with previous studies.

Estimation of (4.7)-(4.8) can be carried out by full information ML. Here, I will estimate the model by the standard "Heckit" two-step procedure, which in a first step estimates (4.8) by Probit ML, and uses the estimated $\hat{\alpha}$ to estimate

$$\ln(Y_i) = X_i\beta + \sigma\rho\phi(X_i\hat{\alpha})/\Phi(X_i\hat{\alpha}) + \epsilon_i \tag{4.13}$$

by OLS in the sample with $D_i = 1$ (second step). The estimating equation (4.13) can be seen as an approximation to moment-based estimation using the condition $\mathrm{E}[Y_i - \mathrm{E}(Y_i|D_i = 1, X_i)|X_i] = 0$, where $\mathrm{E}(Y_i|D_i = 1, X_i)$ is given in (4.10). In particular, the inverse Mills ratio is a first order approximation to the multiplicative correction term in (4.10) (Greene, 1998).

The objects of interest are ATE, IME and EME associated with the policy intervention

of reducing the number of bureaucratic procedures; they can be computed from the parameter estimates of the probit equation and of (4.13) using the formulas provided in the preceding section.

4.4.1 Data

The data is taken from the study by Helpman, Melitz and Rubinstein (2008), which the authors kindly make publicly available on the internet (the data can be downloaded from http://www.economics.harvard.edu/faculty/helpman/Data_Sets_Helpman). It is pooled from different sources, including Feenstra's World Trade Flows, the Penn World Tables and the World Bank's World Development Indicators; and is described in detail in Helpman, Melitz and Rubinstein's (2008) Appendix I. Part of their analysis uses country-level data on regulation costs of firm entry collected by Djankov et al. (2002). Specifically, they create two dummy variables indicating a country-pair having high regulation costs. The first is based on the number of bureaucratic procedures it takes to set up a business in a given country, and the number of days that it takes to complete these procedures. The variable equals one when both countries in the pair are above the median according to these criteria. Similarly, the second dummy equals one when importer and exporter are above the median according to regulation costs as measured as a percentage of countries' GDP. In addition to these two binary variables, I use the sum of the number of procedures required in the importing and in the exporting country of a pair; this is the variable of interest in this application.

Descriptive statistics for the variables used in this analysis are presented in Table 4.2. They correspond to the year 1986 (with the exception of the regulation cost variables, which are from 1999). The dataset consists of 11,978 country-pairs of 106 exporter countries and 114 importer countries. The asymmetry stems from countries serving the whole (sampled) world as exporters. As it is necessary to estimate sets of exporter and importer fixed effects to control for multilateral effects (Anderson and van Wincoop, 2003; Feenstra; 2004), all-world exporters were dropped from the dataset to avoid perfect prediction in the decision-

to-trade equation. As can be seen from the number of observations for the logarithm of bilateral trade, only 6,572 out of 11,978 (or 55%) of the country-pairs engage in trade. To explain trade flows, I broadly follow the specification of Helpman, Melitz and Rubinstein. The regressors include great-circle distance between capitals in log-Kilometers (*Distance*), the gravity equation variable par excellence. To capture geography-related trade costs further, the indicator variables *Landlock* (at least one country in pair is landlocked), *Island* (at least one country in pair is an island), and *Land border* (countries share a common border) are used in the specification. Cultural and historical similarities are proxied by the dummy variables *Legal* (origin of legal systems of the countries are the same), *Language* (countries have common language), *Colonial ties* (one country was/is the other's colony) and *Religion*, a continuous index ranging from 0 to 1, which aggregates the similarity in the composition of Catholics, Protestants and Muslims in the countries. As discussed above, regulation costs are mapped by the indicators *Reg. costs (% GDP)* and *Reg. costs (days & proc.)*, as well as *No. of procedures*.

4.4.2 Estimation results

The estimated coefficients of a two-stage Heckit procedure are reported in Table 4.3. The explanatory variables included a set of importer and exporter fixed effects. Due to collinearity, it was not possible to estimate a separate exporter fixed effect for Chad in neither of the two equations. Thus, there is only one joint exporter fixed effect for South Africa, the base-category country, and Chad.

Despite the slightly different data set and specification, the coefficients in Table 4.3 are very similar to the results of Helpman, Melitz and Rubinstein (2008). Specifically, I would like to focus on the effect of the following policy intervention: cutting back two bureaucratic procedures. Two procedures correspond to about half a standard deviation of the variable, and one can think of the intervention as both importer and exporter country eliminating each one bureaucratic hurdle. The effect of the number of procedures on trade is

large, judging from the results in Table 4.3: Two procedures less is expected to increase the probability of trading by about 12%-points for the average country-pair (here: $X_i\hat{\alpha} = 0.46$); a trading country-pair is expected to increase its trade volume by about 146% ($= \exp(-2 \times -0.45) - 1$) on average in response to such a policy intervention. Of course, such causal interpretations are only valid if all model assumptions hold; in particular, if regressors are independent of errors.

Consider as an example country-pairs for which the average predicted probability of participation is 0.5. Observations with such an average probability include trade from Romania to Bolivia or from Ethiopia to South Korea, which is positive; as well as trade from Romania to Honduras or from Cambodia to South Korea, which is zero. Table 4.4 shows the estimated ATE, EME and IME for such country-pairs. In the first row of Table 4.4 it is assumed that $\sigma\rho = 0$. As discussed previously, the two decompositions (4.1) and (4.3) of the ATE coincide in this case. The model has been estimated using $\ln(y)$ as the dependent variable. To retransform predictions to levels, an estimate of σ^2 is needed. Such an estimate is not directly available because the two-step estimator only gives $\widehat{\sigma\rho}$. Therefore, Duan's (1983) smearing estimator was used to obtain predictions in levels. The total effect of 893 corresponds to an increase in expected trade flows of about 140%. Of the 893, the extensive margin contributes 39%.

But assume first that $\sigma\rho = 0$, and calculate the total effect and its decomposition under this premise (first row in Table 4.4). As discussed previously, the two decompositions (4.1) and (4.3) of the ATE coincide in this case. The model has been estimated using $\ln(y)$ as the dependent variable. To retransform predictions to levels, an estimate of σ^2 is needed. Such an estimate is not directly available because the two-step estimator only gives $\widehat{\sigma\rho}$. Therefore, Duan's (1983) smearing estimator was used to obtain predictions in levels. The total effect of 893 corresponds to an increase in expected trade flows of about 140%. Of the 893, the extensive margin contributes 39%.

The estimated coefficient on the inverse Mills ratio, $\widehat{\sigma\rho}$, is 0.20 with a p-value of 2.4%,

suggesting that there is some dependence between errors. Rows (2)-(5) of Table 4.4 show estimates of the total effect and its decomposition with correlation. Row (2) contains an estimate as may be found in previous literature. It uses the standard decomposition (4.1), and approximates the conditional expectation function $\mathrm{E}(Y_i|D_i = 1, X_i)$ by $\hat{\eta}\exp(X_i\hat{\beta} + \widehat{\sigma\rho}\hat{\lambda}_i)$, where $\hat{\eta}$ is used to denote the smearing estimate of $\mathrm{E}(\exp(\epsilon)|x)$ and $\hat{\lambda}_i$ is the estimated inverse Mills ratio for X_i. Row (4) contains results of the same decomposition but using the exact functional forms of section 4.3.2. It turns out that the difference between using approximate (with inverse Mills ratio) and exact expectations (with multiplicative correction term) is negligible in this application.

Regarding the relative contribution of extensive and intensive country margin to the total effect, the differences between omitting correlation and taking it into account are small in this case (39% vs. 40%). The magnitude of the total effect is also quite close to the one ignoring correlation. Thus, the relative effect necessarily needs to be smaller, since the expectations conditioning on the error correlation are larger than the ones omitting the adjustment factor because $\Phi(\sigma\rho + X_i\alpha)/\Phi(X_i\alpha) > 1$ for positive ρ. The difference is almost 20%-points (139% vs. 122%).

Comparing the results of the conventional decomposition to the one proposed in this paper based on country-types, the shift in the contribution of the margins is clearly visible. The extensive margin contribution decreases by 6%-points going from row (4) to (5), a 15% difference. The reason for this drop is that with positive correlation, the doubly-truncated expectation of the error underlying the new decomposition will always be lower than the error expectation truncated from below at the upper bound of the doubly-truncated expectation which underlies the conventional method. The total effect is the same. While for the approximations in rows (2) and (3) this identity is slightly veiled, it holds precisely when using the exact functional forms in rows (4) and (5).

The overestimation of the extensive margin is not limited to this group of country-pairs, of course. For instance, using country-pairs at the mean $X_i\hat{\alpha}$, the overestimation of the

extensive margin by the standard decomposition is over 20%. Using the estimated value of ρ (= 0.2), and of α_T (= $-2 \times -0.2 = 0.4$), Fig. 4.2 plots the ratio of conventional versus causal EME (cf. Eq. 4.11) over some of the range of the participation probability, $\Phi(X_i\alpha)$. The increase in overestimation is quite steep.

4.5 Discussion

This paper presented a decomposition of average treatment effects in corner solution models into extensive and intensive margins based on the joint distribution of potential outcomes. The new decomposition is a weighted sum of the ATE of subgroups of the population — switchers and participants—, and it differs markedly from the traditional decomposition, which lacks an interesting causal interpretation. This was demonstrated in a numerical example for the Tobit model, and in a substantive application to international trade flows for a generalization of the Tobit model. By relying on very strong distributional assumptions, these models display tractable closed forms which are useful for both illustration and for comparison with previous research.

However, the decomposition of treatment effects presented here is also applicable to semiparametric models. One such class are latent factor structure models (Aakvik, Heckman and Vytlacil, 2005). These models relax the functional form assumptions, but in turn require an exclusion restriction. I.e., an instrumental variable is needed which affects switchers but not participants. An alternative assumption which would have identifying power within linear-index models would be that participants and switchers display (different) index heteroskedasticity, as in Klein and Vella (2009). Both functional form and exclusion restriction assumptions are non-refutable, but their plausibility might differ depending on the application.

While the choice of decomposition matters in general, it makes no difference under the two-part model. The reason for this is the assumed error independence in the class

of two-part models I considered. This assumption is unattractive, but it seems that it is not a necessary ingredient of two-part models. Duan et al. (1984) provide an example of a two-part model with correlated errors where the correlation parameter does not enter the likelihood function. Thus, consistent parameter estimates can still be obtained conveniently by separate probit and linear regressions. In such two-part models, the correct decomposition *would* differ from the traditional — however, neither decomposition could be calculated (nor even the ATE) as the correlation parameter is unavailable. Thus, this hardly seems to make the two-part model more attractive for causal inference.

The causally meaningful decomposition of the ATE does not come free of cost: It requires more assumptions than needed for the ATE alone, as it deals with potential outcomes jointly. As applied work often makes assumptions that go well beyond the required for the decomposition, however, choosing the correct decomposition does seem to be almost free of cost. For instance, all the applied articles cited in the introduction imposed enough structure to point-identify the causal decomposition.

Finally, while this article examined the decomposition into effects at margins in a simple experimental setting where nonparametric identification fails, other experimental settings can be devised which have point-identifying power under weaker conditions. Pre-treatment measurements of Y_i is one such setting. If it can be ensured that individual-specific unobservables do not vary over time ($U_i = U_{it}$ for periods $t = 0, 1$), the decomposition is nonparametrically identified.

References

Aakvik, Arild, James J. Heckman and Edward J. Vytlacil (2005), "Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs", *Journal of Econometrics*, **125**(1), 15-51.

Amemiya, Takeshi (1985), *Advanced Econometrics*, Harvard University Press.

Anderson, James E. and Eric van Wincoop (2003), "Gravity with Gravitas: A Solution to the Border Puzzle", *American Economic Review*, **93**(1), 170-192.

Angrist, Joshua D. (2001), "Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice", *Journal of Business and Economic Statistics* **19**(1), 2-16.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009), *Mostly Harmless Econometrics*, Princeton University Press.

Variables:

Baranga, Thomas (2009), "Unreported Trade Flows and Gravity Equation Estimation", unpublished manuscript, IRPS, UC San Diego.

Cameron, A. Colin and Pravin K. Trivedi (2005), *Microeconometrics*, Cambridge University Press.

Caspi, Avshalom, Bradley R. Entner Wright, Terrie E. Moffitt, Phil A. Silva (1998), "Early Failure in the Labor Market: Childhood and Adolescent Predictors of Unemployment in the Transition to Adulthood" *American Sociological Review* **63**(3), 424-451.

Chen, Songnian (2010), "Non-Parametric Identification and Estimation of Truncated Regression Models", *Review of Economic Studies*, **77**(1), 127-153.

Chiburis, Richard C. (2010), "Semiparametric bounds on treatment effects", *Journal of Econometrics*, **159**(2), 267-275.

Co, Catherine Yap (2010), "Intra- and inter-firm US trade ", *International Review of Economics and Finance*, **19**(2), 260-277.

Cragg, John G. (1971), "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods", *Econometrica*, **39**(5), 829-844.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes and Andrei Shleifer, (2002), "The Regulation of Entry", *Quarterly Journal of Economics*, **117**(1), 1-37.

Dow, William H. and Edward C. Norton (2003), "Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions", *Health Services and Outcomes Research Methodology*, **4**(1), 5-18.

Duan, Naihua (1983), "Smearing estimate: a nonparametric retransformation method", *Journal of the American Statistical Association*, **78**(3), 605-610.

Duan, Naihua, Willard G. Manning, Jr., Carl N. Morris, Joseph P. Newhouse (1983), "A Comparison of Alternative Models for the Demand for Medical Care", *Journal of Business and Economic Statistics*, **1**(2), 115-126.

Duan, Naihua, Willard G. Manning, Jr., Carl N. Morris, Joseph P. Newhouse (1984), "Choosing between the Sample-Selection Model and the Multi-Part Model", *Journal of Business and Economic Statistics*, **2**(3), 283-289.

Engelhardt and Kumar (2007), "Employer matching and 401(k) saving: Evidence from the health and retirement study", *Journal of Public Economics*, **91**(10), 1920-1943.

Fan, Yanqin and Jisong Wu (2010), "Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and its Confidence Sets", *Review of Economic Studies*, **77**(3), 1002-1041.

Feenstra, Robert C. (2004), *Advanced International Trade: Theory and Evidence*, Princeton University Press.

Feenstra, Robert C. (2008), "Gravity Equation", in: *The New Palgrave Dictionary of Economics*, Second Edition, Eds. Steven N. Durlauf and Lawrence E. Blume, Palgrave Macmillan.

Felbermayr, Gabriel J. and Wilhelm Kohler (2006), "Exploring the intensive and extensive margin of world trade", *Review of World Economics 2006*, **142**(4), 642-674.

Greene, William H. (1998), "Sample selection in credit-scoring models", *Japan and the World Economy*, **10**(3), 299-316.

Greene, William H. (2008), *Econometric Analysis*, 6th edition, Prentice Hall.

Hallak, Juan Carlos (2006), "Product quality and the direction of trade", *Journal of International Economics*, **68**(1), 238-365.

Hastings, Justine and Ebonya Washington (2010), "The First of the Month Effect: Consumer Behavior and Store Responses." *American Economic Journal: Economic Policy*, **2**(2), 142-162.

Hay, Joel W. and Randall J. Olsen (1984), "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care", *Journal of Business and Economic Statistics*, **2**(3), 279-282.

Heckman, James J. (1979), "Sample Selection Bias as a Specification Error", *Econometrica*, **47**(1), 153-161.

Helpman, Elhanan, Marc Melitz and Yona Rubinstein (2008), "Estimating Trade Flows: Trading Partners and Trading Volumes", *Quarterly Journal of Economics*, **123**(2), 441-487.

Joulfaian, David (2000), "Corporate Income Tax Evasion and Managerial Preferences", *Review of Economics and Statistics*, **82**(4), 698-701.

Kenkel, Donald S. (1991), "Health Behavior, Health Knowledge, and Schooling", *Journal of Political Economy*, **99**(2), 287-305.

Klein, Roger and Francis Vella (2009), "A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroscedasticity", *Journal of Applied Econometrics*, **24**(5), 735-762.

Liu, Xuepeng (2009), "GATT/WTO Promotes Trade Strongly: Sample Selection and Model Specification", *Review of International Economics*, **17**(3), 428-446.

Manski, Charles F. (1997), "Monotone treatment response", *Econometrica*, **65**(6), 1311-1334.

Manski, Charles F. (2003), *Identification for Prediction and Decision*, Harvard University Press.

McDonald, John F. and Moffitt, Robert A. (1980), "The Uses of Tobit Analysis", *Review of Economics and Statistics*, **62**(2), 318-321.

Moffitt, Robert A. (1982), "The Tobit Model, Hours of Work and Institutional Constraints", *Review of Economics and Statistics*, **64**(3), 510-515.

Santos Silva, João M.C. and Silvana Tenreyro (2006), "The Log of Gravity", *Review of Economics and Statistics*, **88**(4), 641-658.

Sousa-Poza, Alfonso and Alexandre Ziegler (2003), "Asymmetric information about workers' productivity as a cause for inefficient long working hours", *Labor Economics*, **10**(6), 727-747.

Terza, Joseph V. (1998), Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects, *Journal of Econometrics*, **84**(1), 129-154.

Tobin, James (1958), "Estimation of Relationships for Limited Dependent Variables", *Econometrica*, **26**(1), 24-36.

Trejo, Stephen S. (1993), "Overtime Pay, Overtime Hours, and Labor Unions", *Journal of Labor Economics*, **11**(2), 253-278.

Vytlacil, Edward (2002), "Independence, Monotonicity and Latent Index Models: An Equivalence Result", *Econometrica*, **77**(1), 331-341.

Wooldridge, Jeffrey M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

Table 4.1: Features of participants and switchers in the Tobit model

	Participants $(Y_{0i} > 0, Y_{1i} > 0)$	Switchers $(Y_{0i} = 0, Y_{1i} > 0)$
U_i	$(-\beta_0, \infty)$	$(-\beta_0 - \beta_1, -\beta_0)$
$\Pr(Y_{1i}, Y_{0i})$	Φ_0	$\Phi_1 - \Phi_0$
$Y_{1i} - Y_{0i}$	β_1	$\beta_0 + \beta_1 + U_i$
$\mathrm{E}(Y_{1i} - Y_{0i})$	β_1	$\beta_0 + \beta_1 + \sigma \frac{\phi_1 - \phi_0}{\Phi_1 - \Phi_0}$

Notes: $\Phi_T = \Phi(\beta_0 + \beta_1 T)$, $\phi_T = \phi(\beta_0 + \beta_1 T)$, for $T = 0, 1$. $\Phi(\cdot)$ is the standard normal cdf, $\phi(\cdot)$ the standard normal pdf. The Tobit model in this table has the latent variable $Y_i^* = \beta_0 + \beta_1 T_i + U_i$, with $U_i | T_i \sim N(0, \sigma^2)$ and $\beta_1 > 0$.

Table 4.2: Summary statistics

Variable	Mean	Std. Dev.	Min.	Max.	No. of obs.
Bilateral trade	79,804.40	991,081.28	0	74,558,336	11,978
Log of Bilateral Trade	8.33	3.04	1.61	18.13	6,572
Distance	4.17	0.8	0.3	5.66	11,978
Land border	0.02	0.15	0	1	11,978
Island	0.17	0.37	0	1	11,978
Landlock	0.36	0.48	0	1	11,978
Legal	0.36	0.48	0	1	11,978
Language	0.26	0.44	0	1	11,978
Colonial ties	0.01	0.09	0	1	11,978
Religion	0.17	0.25	0	0.99	11,978
No. of procedures	19.59	4.86	4	36	11,978
Reg. costs high (%GDP)	0.33	0.47	0	1	11,978
Reg. costs high (days & proc.)	0.12	0.33	0	1	11,978

Source: Data are from Helpman, Melitz and Rubinstein (2008), available online. See text Section 4.1.

Table 4.3: Estimated coefficients — Two-equations model of bilateral trade

Regression	$\Pr(d=1\|x)$	$\mathrm{E}(\ln y\|y>0,x)$
Method	ML (Probit)	OLS
	(1)	(2)
Distance	-0.62**	-1.22**
	(0.03)	(0.04)
Land border	-0.16	0.67**
	(0.13)	(0.14)
Island	-0.54*	-0.44
	(0.24)	(0.25)
Landlock	-0.14	-0.42*
	(0.12)	(0.17)
Legal	0.15**	0.55**
	(0.04)	(0.06)
Language	0.32**	0.19**
	(0.06)	(0.07)
Colonial ties	-0.02	0.89**
	(0.35)	(0.19)
Religion	0.33**	0.32**
	(0.09)	(0.11)
No. of procedures	-0.20**	-0.45**
	(0.03)	(0.05)
Reg. costs high (%GDP)	-0.27**	-0.13
	(0.08)	(0.09)
Reg. costs high (days & proc.)	-0.16*	-0.25*
	(0.07)	(0.11)
Inv. Mills ratio		0.20*
		(0.09)
R^2	0.57	0.69
$\log L$	-3,580	-12,760
Observations	11,978	6,572

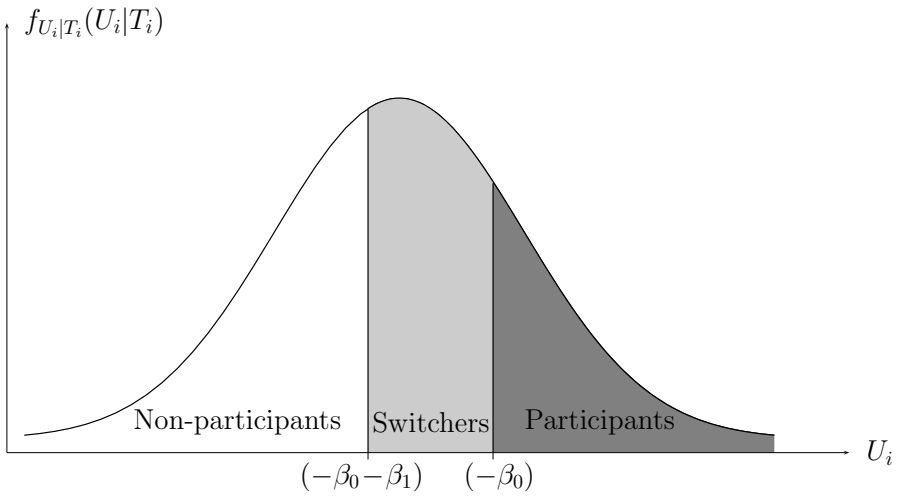
Notes: Robust standard errors in parentheses. $*$ and $**$ denote statistical significance on the 5% and 1% level. Additional regressors include a constant term and a complete set of importer fixed effects and of exporter fixed effects.

Table 4.4: Total trade effects and decomposition into country margins

Decomposition	TE	TE/E(y|x)	EME (% of TE)	IME (% of TE)
(1) No error correlation	893	139%	347	546
			(39%)	(61%)
(2) Conventional decomposition, approx.	888	122%	353	535
			(40%)	(60%)
(3) Decomposition by country-type, approx.	892	123%	308	584
			(35%)	(65%)
(4) Conventional decomposition, exact	879	122%	349	530
			(40%)	(60%)
(5) Decomposition by country-type, exact	879	122%	302	577
			(34%)	(66%)

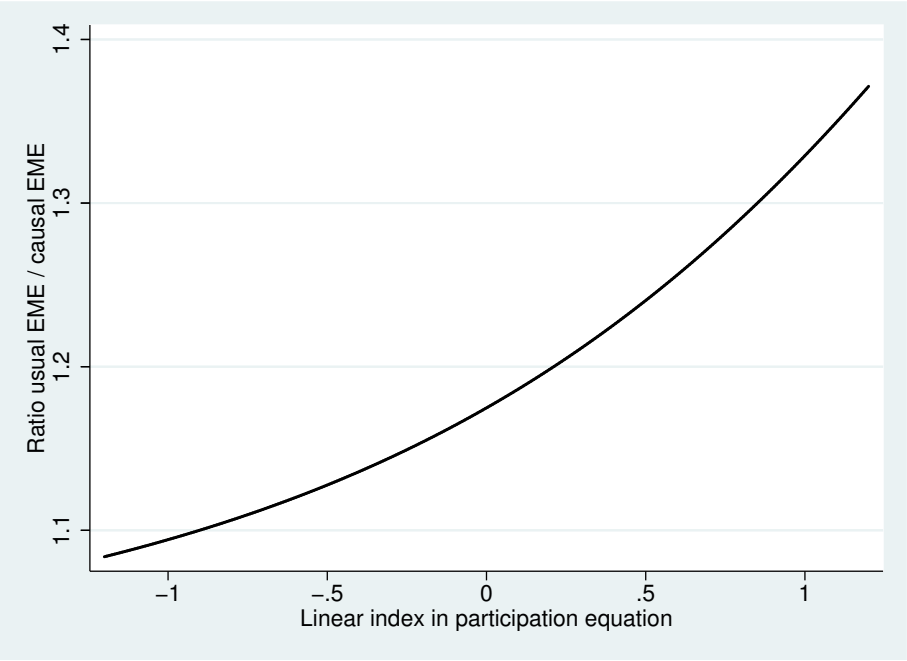
Notes: Own calculations based on results from Table 4.3. TE stands for Total Effect, EME for Extensive Margin Effect, and IME for Intensive Margin Effect. Formulas are discussed in Section 3.2. Effects are for a country-pair with $X\hat{\alpha} = -0.2$ and $X\hat{\beta} = 5.32$.

Figure 4.1: Population groups by U_i in the Tobit model



Notes: The Tobit model in this figure has the latent variable $Y_i^* = \beta_0 + \beta_1 T_i + U_i$, with $U_i|T_i \sim N(0, \sigma^2)$ and $\beta_1 > 0$.

Figure 4.2: Overestimation of extensive margin effect (EME) in estimated trade model



Notes: Own calculations based on results from Table 4.3. "Base probability of participation" means participation probability without treatment (reduction in "No. of proc."). Base probability of participation plotted over range 0.1–0.9. "Ratio usual EME / causal EME" is $\widehat{\widetilde{\text{EME}}}/\text{EME}$ as in Eq. (4.11), calculated for estimated values of $\widehat{\sigma\rho}$ (= 0.2) and $\hat{\alpha}_T$ ($= -2 \times -0.2 = 0.4$).

Chapter 5

Consistent estimation of the fixed effects ordered logit model

This chapter is joint work with Gregori Baetschmann and Rainer Winkelmann.

Acknowledgements: We thank Paul Frijters, Arie Kapteyn and participants of the 2011 Engelberg Workshop in Labor Economics for very valuable comments on an earlier draft.

5.1 Introduction

Economists' use of panel data has been increasing steadily over the past years. As a key advantage, such data potentially allow to solve the endogeneity problem arising from correlated time-invariant unobserved individual-specific effects. However, while implementing corresponding estimators is straightforward if the model is linear, no generally valid method exists for non-linear models.

One important application of non-linear models arises in the context of responses that are coded on a discrete and ordinal scale. Such scales are prominent in many household surveys, where they provide information on subjective assessments, judgments, or expectations. Examples are an individuals' satisfaction (with one's job, life in general, etc.) or expectations about the future (of the economy, of one's income, etc.). There are good reasons for using such subjective evaluations in empirical research, for example because they substitute for objective information that is not collected, or because the subjective responses are of their own interest. For instance, subjective health status might be more closely tied to certain behavioral responses than actual health.

The most popular regression-type models for such dependent variables are the ordered probit model and, in particular, the ordered logit model. With cross-section data, these parametric models are very easy to use and to estimate by maximum likelihood. However, extensions to a panel data context are complex and far from obvious. Unlike in the linear model, no simple transformation (such as first-differencing or within-transformation) is available that would purge the ordered response models from the individual-specific fixed effects.

While the situation is hopeless for the ordered probit model, it is more favorable for the ordered logit model, where it has been recognized early on that the estimation problem can be simplified to that of a binary logit model for which a fixed effects estimator exists, by collapsing the J categorical responses into two classes (e.g. Winkelmann and Winkelmann, 1998). The binary logit fixed effects estimator, due to Chamberlain (1980), uses the fact

that conditioning the individual likelihood contribution on the sum of the outcome over time provides an expression which is independent of the fixed effects. The effect of the time-varying regressors can then be estimated by conditional maximum likelihood (CML).

Other popular estimators for the fixed effects ordered logit model proposed in the literature are also built around the same idea of reducing the ordered model to a binary one, but aim at improving over a simple dichotomization by exploiting additional information available in the data. One approach is to estimate fixed effects logits with every possible dichotomizing cutoff point, and then combine the resulting estimates by minimum distance estimation (Das and van Soest, 1999).

A second approach is to dichotomize every individual separately, at some sort of 'optimal' or 'efficient' cutoff point (Ferrer-i-Carbonell and Frijters, 2004). The most popular variant dichotomizes the dependent variable at the individual mean, which ensures that every individual displaying some time variation in the outcome is included in the estimation. Such fixed effects ordered logit models have been used frequently in the literature. Recent applications to health economics include Jones and Schurer (2009), and Frijters, Haisken-DeNew and Shields (2004 a, b); additions to the satisfaction literature comprise Kassenboehmer and Haisken-DeNew (2009), Booth and van Ours (2008), D'Addio, Eriksson and Frijters (2007), and Frijters, Haisken-DeNew and Shields (2004).

In this article we propose a new consistent estimator for the ordered logit model with fixed effects. We then compare the existing and the new estimators in Monte Carlo simulations. There are two important findings which have implications for future applied research based on panel ordered logit models. First, we observe that individual-specific dichotomized estimators are biased in finite samples. Worse, the bias does not vanish as simulations with increased sample size are considered. We provide reasons for this observation, and show that, in general, these estimators are inconsistent. The problem is that by choosing the cutoff point based on the outcome, they produce a form of endogeneity. Second, we provide evidence on the good finite sample performance of Das and van Soest's

(1999) estimator and our new estimator. In contrast to Das and van Soest's (1991), the new estimator remains unbiased even in very small samples. Moreover, it can be easily implemented using existing software for CML logit estimation.

The paper proceeds as follows. Section 5.2 presents the different estimators for the fixed effects ordered logit models. Then, we explain our Monte Carlo simulation setup and discuss its results (Section 5.3), followed by an application of the estimators to data from the German Socioeconomic Panel which studies the effect of unemployment on life satisfaction (Section 5.4). Section 5.5 concludes.

5.2 Estimators for the FE ordered logit model

5.2.1 The FE ordered logit model

The fixed effects ordered logit model relates the latent variable y_{it}^* for individual i at time t to a linear index of observable characteristics x_{it} and unobservable characteristics α_i and ε_{it}:

$$y_{it}^* = x_{it}'\beta + \alpha_i + \varepsilon_{it}, \quad i = 1, \ldots, N \quad t = 1, \ldots, T \tag{5.1}$$

The time-invariant part of the unobservables, α_i, can be statistically dependent of x_{it}. In this case, one can either make an assumption regarding the joint distribution of α_i and x_{it}, or else treat α_i as a fixed effect. This paper considers estimation under the fixed effects approach.

The latent variable is tied to the (observed) ordered variable y_{it} by the observation rule:

$$y_{it} = k \quad \text{if} \quad \tau_k < y_{it}^* \leq \tau_{k+1}, \quad k = 1, \ldots, K$$

where thresholds τ are assumed to be strictly increasing ($\tau_k < \tau_{k+1} \quad \forall k$) and $\tau_1 = -\infty$, $\tau_{K+1} = \infty$. It is possible to formulate the model more generally with individual-specific thresholds (Ferrer-i-Carbonell and Frijters, 2004):

$$y_{it} = k \quad \text{if} \quad \tau_{ik} < y_{it}^* \leq \tau_{ik+1}, \quad k = 1, \ldots, K \tag{5.2}$$

The distributional assumption completing the specification of the fixed effects ordered logit model is that conditionally on x_{it} and α_i, ε_{it} are IID standard logistically. I.e., if $F(\cdot)$ denotes the cdf of ε_{it}

$$F(\varepsilon_{it}|x_{it},\alpha_i) = F(\varepsilon_{it}) = \frac{1}{1+\exp(-\varepsilon_{it})} \equiv \Lambda(\varepsilon_{it}) \tag{5.3}$$

Hence, the probability of observing outcome k for individual i at time t using (5.1), (5.2), (5.3) is

$$\Pr(y_{it}=k|x_{it},\alpha_i) = \Lambda(\tau_{ik+1} - x'_{it}\beta - \alpha_i) - \Lambda(\tau_{ik} - x'_{it}\beta - \alpha_i) \tag{5.4}$$

which depends not only on β, but also on α_i and τ_{ik},τ_{ik+1}.

There are two problems with Maximum Likelihood (ML) estimation based on expression (5.4). The first is a problem of identification: τ_{ik} cannot be distinguished from α_i; only $\tau_{ik} - \alpha_i \equiv \alpha_{ik}$ is identified and can thus, in principle, be estimated consistently for $T \to \infty$. The second problem arises, since in most applications, T must be treated as fixed and relatively small. But under fixed-T asymptotics even α_{ik} cannot be estimated consistently, due to the incidental parameter problem (see, for instance, Lancaster, 2000). This does have consequences for estimation of β – the bias in α_{ik} contaminates $\hat{\beta}$. In short panels, the resulting bias in $\hat{\beta}$ can be substantial (Greene, 2004).

We next consider different approaches to estimate β consistently. They all use the same idea of collapsing y_{it} into a binary variable ($y_{it} \geq k$ and $y_{it} < k$) and then applying the sufficient statistic suggested by Chamberlain (1980) to construct a CML estimator.

5.2.2 Chamberlain's CML estimator for the dichotomized ordered logit model

Let d_{it}^k denote the binary dependent variable that results from dichotomizing the ordered variable at the cutoff point k: $d_{it}^k = \mathbb{1}(y_{it} > k)$. By construction, $P(d_{it}^k = 0) = P(y_{it} \leq k) = \Lambda(\tau_{ik+1} - x'_{it}\beta - \alpha_i)$, and $P(d_{it}^k = 1) = P(y_{it} > k) = 1 - \Lambda(\tau_{ik+1} - x'_{it}\beta - \alpha_i)$. Now consider

the joint probability of observing $d_i = (d_{i1}^k, \ldots, d_{iT}^k) = (j_{i1}, \ldots, j_{iT})$ with $j_{it} \in \{0, 1\}$. The sum of all the individual outcomes over time is a sufficient statistic for α_i as

$$\mathcal{P}_i^k(\beta) \equiv \Pr\left(d_i^k = j_i \middle| \sum_{t=1}^{T} d_{it}^k = a_i\right) = \frac{\exp(j_i' x_i \beta)}{\sum_{j \in B_i} \exp(j' x_i \beta)} \tag{5.5}$$

does not depend on α_i and the thresholds. In (5.5), $j_i = (j_{i1}, \ldots, j_{iT})$, x_i is the $(T \times L)$-matrix with tth row equal to x_{it}, L is the number of regressors and $a_i = \sum_{t=1}^{T} j_{it}$. The sum in the denominator goes over all vectors j which are elements of the set B_i

$$B_i = \left\{ j \in \{0, 1\}^T \middle| \sum_{t=1}^{T} j_t = a_i \right\},$$

i.e., over all possible vectors of length T which have as many elements equal to 1 as the actual outcome of individual i (a_i). The number of j-vectors in B_i, and therefore of terms in the sum in the denominator of (5.5), is $\binom{T}{a_i} = \frac{T!}{a_i!(T-a_i)!}$.

Chamberlain (1980) shows that maximizing the conditional likelihood

$$\log \mathcal{L}^k(b) = \sum_{i=1}^{N} \log \mathcal{P}_i^k(b) \tag{5.6}$$

gives a consistent estimate for β (subject to mild regularity conditions on the distribution of α_i, cf. Andersen, 1970). I.e. the score —the gradient of the log-likelihood with respect to β— converges to zero when evaluated at the true β:

$$\text{plim} \frac{1}{N} \sum_i s_i^k(\beta) = 0, \tag{5.7}$$

where

$$s_i^k(b) = \frac{\partial \ln \mathcal{P}_i^k(b)}{\partial b} = x_i' \left(d_i^k - \sum_{j \in B_i} j \frac{\exp(j' x_i b)}{\sum_{l \in B_i} \exp(l' x_i b)} \right) \tag{5.8}$$

The reason why (5.7) holds is that for every i, conditional on x_i, the expectation of the term in parentheses in (5.8) is zero as it defines a conditional expectation residual.

Note that conditioning on a_i causes all time-invariant elements in (5.4) to cancel. I.e., not only α_i and τ_{ik}, τ_{ik+1} are not estimated, but also elements of the β vector corresponding to observables that do not change over time. Also, individuals with constant d_{it}^k do not

contribute to the conditional likelihood function, as $P(d_i^k = 1|\sum_{t=1}^T d_{it}^k = T) = P(d_i^k = 0|\sum_{t=1}^T d_{it}^k = 0) = 1$.

The Hessian is

$$H_i^k(b) = \frac{\partial^2 \ln \mathcal{P}_i^k(b)}{(\partial b)(\partial b)'} = -\sum_{j \in B_i} \frac{\exp(j' x_i b)}{\sum_{l \in B_i} \exp(l' x_i b)} \times \left(x_i' j - \sum_{m \in B_i} \frac{\exp(m' x_i b)}{\sum_{l \in B_i} \exp(l' x_i b)} m' x_i \right) \left(x_i' j - \sum_{m \in B_i} \frac{\exp(m' x_i b)}{\sum_{l \in B_i} \exp(l' x_i b)} m' x_i \right)' \quad (5.9)$$

5.2.3 Combining all possible dichotomizations: Das and van Soest's (1999) two-step estimation, and a new approach

The estimator of β based on (5.6), say $\hat{\beta}^k$, does not use all the variation in the ordered dependent variable y_{it}, as individuals for which either $y_{it} < k$ or $y_{it} \geq k$ for every t do not contribute to the log-likelihood. Since every $\hat{\beta}^k$ for $k = 2, \ldots, K$ provides a consistent estimator of β, and every individual with some variation in y_{it} will contribute to at least one log-likelihood $\mathcal{L}^k(b)$, one can perform CML estimation on all possible $K-1$ dichotimizations and then, in a second step, combine the resulting estimates. The efficient combination will weight the $\hat{\beta}^k$ by the inverse of their variance (Das and van Soest, 1999):

$$\hat{\beta}^{DvS} = \arg\min_b (\hat{\beta}^{2\prime} - b', \ldots, \hat{\beta}^{K\prime} - b') \Omega^{-1} (\hat{\beta}^{2\prime} - b', \ldots, \hat{\beta}^{K\prime} - b')' \quad (5.10)$$

The variance Ω has entries ω_{gh}, $g = 2, \ldots, K$, $h = 2, \ldots, K$, such that

$$\omega_{gh} = \left[\mathrm{E}\left(\frac{\partial \log \mathcal{P}^g}{\partial b} \right) \left(\frac{\partial \log \mathcal{P}^g}{\partial b} \right)' \right]^{-1} \left[\mathrm{E}\left(\frac{\partial \log \mathcal{P}^g}{\partial b} \right) \left(\frac{\partial \log \mathcal{P}^h}{\partial b} \right)' \right] \left[\mathrm{E}\left(\frac{\partial \log \mathcal{P}_i^h}{\partial b} \right) \left(\frac{\partial \log \mathcal{P}^h}{\partial b} \right)' \right]^{-1}$$

evaluated at $b = \beta$. In practice, the unknown variance Ω is replaced by an estimate $\hat{\Omega}$ which is evaluated at $\hat{\beta}^k$, $k = 2, \ldots, K$. The solution to (5.10) is

$$\hat{\beta}^{DvS} = \left(H' \hat{\Omega}^{-1} H \right)^{-1} H' \hat{\Omega}^{-1} (\hat{\beta}^{2\prime}, \ldots, \hat{\beta}^{K\prime})'$$

where H is the matrix of K-1 stacked identity matrices of dimension L (the size of each $\hat{\beta}^k$). An estimate of the variance of the estimator can be obtained as

$$\widehat{\text{Var}}(\hat{\beta}^{DvS}) = \left(H'\hat{\Omega}^{-1}H\right)^{-1}$$

Because $\hat{\beta}^{DvS}$ is a linear combination of consistent estimators, it is itself consistent. Ferrer-i-Carbonell and Frijters (2004) discuss some small sample issues which might affect the performance of $\hat{\beta}^{DvS}$. For instance, one concern is that $\hat{\Omega}$ might be estimated very imprecisely when for some g and h there are only few observations with nonzero contributions to $\hat{\omega}_{gh}$. This is the case when there is only a small overlap between the samples contributing to the CML logit estimator dichotomized at g and the one dichotomized at h.

Thus, we propose an alternative to this two-step combination of all possible dichotomizations which avoids such problems by estimating all dichotomizations jointly subject to the restriction $\beta^k = \beta \ \forall k = 2, \ldots, K$. Hence, the sample (quasi-) log-likelihood of this restricted CML estimator is

$$\log \mathcal{L}(b) = \sum_{k=2}^{K} \log \mathcal{L}^k(b) \tag{5.11}$$

The score of this estimator is the sum of the scores of the CML logit estimators. Since these are consistent, they converge to zero in probability. It follows that the probability limit of the score of the restricted CML estimator is zero as well, establishing its consistency:

$$\text{plim} \sum_{k=2}^{K} \frac{1}{N(K-1)} \sum_{i=1}^{N} s_i^k(\beta) = \text{plim} \frac{1}{N} \sum_i s_i^2(\beta) + \ldots + \text{plim} \frac{1}{N} \sum_i s_i^K(\beta) = 0, \tag{5.12}$$

Since some individuals contribute to several terms in the log-likelihood this creates dependence between these terms, invalidating the usual estimate of the estimator variance based on the information matrix equality. Instead, a sandwich variance estimator (White, 1982) should be used. We propose using the cluster-robust variance estimator which allows for arbitrary correlation within the various contributions of any individual:

$$\widehat{\text{Var}}(\hat{\beta}) = \left(\sum_{i=1}^{N} \hat{h}_i\right)^{-1} \left(\sum_{i=1}^{N} \hat{s}_i \hat{s}_i'\right)^{-1} \left(\sum_{i=1}^{N} \hat{h}_i\right)^{-1}$$

where $\hat{s}_i$ are the stacked CML scores of individual i evaluated at $\hat{\beta}$, $\hat{s}_i = (\hat{s}_i^{2\prime}, \ldots, \hat{s}_i^{K\prime})'$, and $\hat{h}_i$ is the matrix of derivatives of s_i with respect to β evaluated at $\hat{\beta}$.

We will refer to this estimator as the BUC estimator. The acronym stands for "Blow-Up and Cluster" which describes the way of implementing this estimator using the CML estimator: Replace every observation in the sample by $K-1$ copies of itself ("blow-up" the sample size), and dichotomize every $K-1$ copy of the individual at a different cutoff point. Estimate CML logit using the entire sample; these are the BUC estimates. Cluster standard errors at the individual level. This implementation requires but a few lines of code in standard econometric software (cf. Appendix A, which contains code for implementation in Stata).

5.2.4 Endogenous dichotomization: Ferrer-i-Carbonell and Frijters (2004) and related approaches

The previous approaches used all possible dichotomizations. Ferrer-i-Carbonell and Frijters (2004) proposed an estimator which chooses dichotomizations separately for every individual. The (quasi-) log-likelihood for their estimator can be written as

$$\log \mathcal{L}^{FF}(b) = \sum_{i=1}^{N} \sum_{k=2}^{K} w_i^k \log \mathcal{P}_i^k(b), \quad w_i^k \in 0,1, \quad \sum_{k=2}^{K} w_i^k = 1 \tag{5.13}$$

This objective function is maximized with respect to b after choosing the cutoff point at which to dichotomize each y_i, i.e. after deciding which one of the individual's weight vectors w_i^k is equal to 1.

Ferrer-i-Carbonell and Frijters' (2004) approach here is to calculate for every individual all Hessian matrices under different cutoff points and choosing the smallest:

$$w_i^k = 1 \quad \text{if} \quad k = \arg\min_{\kappa} \frac{\partial^2 \log \mathcal{P}_i^{\kappa}(b)}{(\partial b)(\partial b)'}\Big|_{b=\beta}$$

In practice, the Hessian is evaluated at $\hat{\beta}$, where $\hat{\beta}$ is a preliminary consistent estimator. Since for every possible dichotomization the choice falls on the cutoff point leading

to the smallest Hessian, this rule should yield the estimator of (5.13) with minimal variance. Other, simpler rules for choosing w_i^k for (5.13) have been used, trading efficiency for computational ease. In fact, the standard way in which this estimator is implemented in the applied literature is by choosing the dichotomizing cutoff point as the mean of the dependent variable:

$$w_i^k = 1 \quad \text{if} \quad k = \text{ceil}\left(T^{-1}\sum_t y_{it}\right)$$

where ceil(z) stands for rounding z up to the nearest integer. This ensures that every individual with time-variation in y_i will be part of the estimating sample. Studies using both rules report little difference in estimates and standard errors, which has led to the view that this way of choosing w_i^k is an approximation to Ferrer-i-Carbonell and Frijters' (2004) . An alternative is using the median instead of the mean as a rule to define the individual dichotomization.

Thus, all these procedures choose the dichotomizing cutoff point endogenously, since it depends on y_i. This is obviously problematic and we show in Appendix B that these estimators are, in general, inconsistent. Here we provide some intuition for this result using the mean-cutoff estimator as an example; similar arguments hold for the other estimators.

The problem is not, as one might suspect, that the cutoffs vary between individuals *per se*. For instance, if the variation of the cutoffs between individuals was random, the resulting estimator would be consistent: the score would be a sum of scores of CML logit estimators, much like the BUC estimator (but with $K-1$ times less observations as each individual would contribute only to exactly one CML logit estimator). I.e., in terms of (5.8), for every random individual-specific cutoff, the resulting vectors d_i converge to their respective conditional expectation, yielding an expected score of zero at the limit.

The real problem lies in the endogeneity of the cutoff. For the mean estimator, $d_{it}^{\text{Mn}} = 1$ if and only if $y_{it} \geq T^{-1}\sum_t y_{it}$. Thus, y_{it} itself is part of the cutoff and the probability

$\Pr(d_{it}^{\text{Mn}} = 1)$ can be written as

$$\Pr(d_{it}^{\text{Mn}} = 1) = \Pr\left(y_{it} \geq \frac{1}{T}\sum_{t} y_{it}\right) = \Pr\left(y_{it} \geq \frac{1}{T-1}\sum_{s \neq t} y_{is}\right)$$

The expression after the first equality makes clear that for any t, y_{it} is on both sides of the inequality sign. Solving for y_{it} shows that the probability $\Pr(d_{it} = 1)$ is equal to the probability that the outcome in t is greater than the average outcome in the remaining periods. In general, this is a different dichotomizing cutoff point within the same individual for every period. Thus, although the researcher is setting an individual-specific cutoff, say k, the endogenous way in which this cutoff is chosen implies that it is equivalent to choosing different cutoff points for the same individual. With endogenous cutoffs the conditional distribution of d_i can be shown to differ from the CML terms, and the score of these estimators will, in general, not converge to zero.

5.3 Monte Carlo simulations

We compare the performance of the estimators discussed in the previous section in finite samples using Monte Carlo simulations. To the best of our knowledge, this is the first investigation of these estimators in a Monte Carlo study. The aim is to assess the small sample biases and efficiency across different data generating processes.

5.3.1 Experimental design

The setup of the Monte Carlo experiment is as follows. The data generating process (DGP) for the latent variable is

$$y_{it}^{*} = \beta_x x_{i,t} + \beta_d d_{i,t} + \alpha_i + \varepsilon_{it},$$

and we set $\beta_x = 1$, $\beta_d = 1$. The regressor x is continuous, while d is binary. We follow Greene (2004) in specifying the fixed effects as

$$\alpha_i = \sqrt{T}\bar{x}_i + \sqrt{T}\bar{u}_i, \quad \bar{x}_i = T^{-1}\sum_{t} x_{it}, \quad \bar{u}_i = T^{-1}\sum_{t} u_{it}, \quad u_{it} \sim N(0,1)$$

For the simulations, we use fixed (not individual-specific) thresholds:

$$y_{it} = k \quad \text{if} \quad \tau_k < y_{it}^* \leq \tau_{k+1}, \quad k = 1, \ldots, K$$

Finally, ε_{it} is sampled from a logistic distribution as in (5.3).

The baseline DGP is a balanced panel of N=500 individuals observed for T=4 periods. The continuous regressor x is distributed as standard normal, the binary regressor's probability of a 1 is 50%. The latent variable is discretized into K=5 categories, choosing the thresholds to yield the marginal distribution depicted in the upper left graph in Fig. 5.1. We call this distribution of y "skewed".

The baseline DGP is modified in a number of dimensions, which can be broadly classified into two experiments. First, different kinds of asymptotics are considered by increasing N, T and K. Second, the influence of the data distribution is explored by sampling from different distributions from the regressors, and by shifting the thresholds to yield different marginal distributions for y_{it}. In the following section, we comment on selected results from these experiments. A supplementary appendix containing full simulation output from a comprehensive exploratory study is available from the authors on request.

5.3.2 Results

Table 5.1 contains results for the baseline scenario. Columns contain mean and standard deviation of estimated coefficients (labeled M and SD), as well as the mean of standard errors (labeled SE) corresponding to x (first three columns) and d (last three columns). Every row gives these results for a different estimator. All entries have been rounded to two decimal places.

The first row, named DvS, contains results for the two-step estimator of Das and van Soest (1999). With means of 0.99 for $\hat{\beta}_x$ and 1.00 for $\hat{\beta}_d$ DvS is virtually unbiased. The BUC estimator, whose results are displayed in the second row, produces unbiased results, too. There is almost no perceivable difference in efficiency between the two estimators.

Estimation of the coefficient corresponding to the binary variable is less precise than that of the continuous regressor — its standard deviation is around 60% higher.

The next three rows contain results for Ferrer-i-Carbonell and Frijters' (2004) estimator (named FF), as well as for the variants dichotomizing at the individual mean (labeled Mean) and at the individual median (labeled Median). These three estimators display standard deviations of the same size as BUC's and DvS'. However, their mean shows a clear downward bias, ranging from 8% for FF to 5% for Median. With a standard deviation of 0.07 (0.12) and 1,000 replications, the margin of error at 99% confidence for these biases is less than 0.6% (1%).

The last four rows contain results for CML logit estimators dichotomized at the categories 2 to 5 (named '$y \geq 2$' to '$y \geq 5$'). As DvS and BUC, these estimators show little finite sample bias. The standard deviations are at best about 30% larger than BUC's — this corresponds to cases where the dichotomized dependent variable has a distribution which is as balanced as possible. For '$y \geq 2$' and '$y \geq 3$' the percentage of zeros is 40% and 70%. For '$y \geq 5$' this percentage is 95%, and the standard deviation of the estimator is more than double that of BUC.

Comparing columns containing the standard deviations of the estimators (SD) with columns containing average standard errors (SE) shows that standard errors are estimated satisfactorily in all cases.

Taken together, the results of Table 5.1 contain two important findings. First, there is no evidence that finite sample issues affect the DvS estimator. All estimators exploiting more information in the data than CML logit estimators with fixed cutoffs are, indeed, more efficient than them. However, they all display about the same standard deviation. Second, estimators based on endogenous dichotomizing cutoff points are all biased in this setup.

Next, we want to check whether these results can be generalized to other settings. We start by conducting asymptotic exercises to explore under which conditions the biases of

FF, Mean and Median can be expected to vanish. The results are reported in Table 5.2.

The first panel of Table 5.2 ('Baseline scenario'), consisting of the first four columns, copies the results from Table 5.1 for easier comparability. Columns with averages of standard errors (SE) were dropped to avoid clutter; we found that results for SE were similar to Table 5.1's for all DGPs considered in this paper. In the second panel ('N=1,000', the next four columns) the effect of increasing sample size with fixed T is considered. As expected by the ratio $\sqrt{500}/\sqrt{1,000}$ the standard deviation falls by 30% for all estimators. As before, DvS and BUC are unbiased. However, FF, Mean and Median estimators remain biased. Indeed, their bias is essentially the same with 1,000 individuals as with 500. This suggests that these are not small sample biases, but that they can be attributed entirely to these estimators' inconsistency.

A different asymptotic experiment holds N fixed and increases the number of time periods. Based on the discussion of the inconsistency of estimators with endogenous dichotomization, we would expect this to have an attenuating impact on their biases: As T increases, the contribution of any y_{it} to the endogenous cutoff (a function of all y_{it} of an individual) decreases. If its contribution was zero, the cutoff would be exogenous. For instance, this is particularly transparent for the mean estimator. In the probability $\Pr(d_{it}^{\text{Mn}} = 1) = \Pr\left(y_{it} > \frac{1}{T-1}\sum_{s\neq t} y_{is}\right)$, the threshold consisting of the average y_{is}, $s \neq t$, becomes less variable for different t as T increases.

The results for this experiment are reported in the next panel, labeled 'T=8', where the number of time periods in the simulations were duplicated from T=4 to T=8. The decrease in the standard deviations relative to the Baseline scenario are of the same magnitude as in the experiment with N=1,000 because here $\sqrt{4}/\sqrt{8} = 1/\sqrt{2}$ as before. Clearly, the biases of FF, Mean and Median are reduced, consistent with our expectation.

A last kind of informal asymptotic experiment which can be conceived is increasing the number of categories. In the limit, the observed variable would be equal to the continuous latent variable. We increase the number of categories from K=5 to K=10, setting the

marginal distribution to the one displayed in the lower right panel of Fig. 5.1. While this distribution is skewed, too, it is of course not exactly the same as in the baseline case. The results are displayed in the fourth panel in Table 5.2, labeled 'K=10'. There are now 5 additional CML logit estimators ($y \geq 6$ to $y \geq 10$), but for the sake of brevity we omit results for these. While Dvs and BUC are almost invariant to the increase in the number of ordered categories, the three estimators based on endogenous dichotomization worsen in terms of bias. This, too, is to be expected. With increasing K and fixed T, the sensitivity of endogenous cutoffs to a particular y_{it} will increase in general. For the mean estimator, for instance, the variance in the mean y_{is}, $s \neq t$ increases with K. It is interesting to note that the median estimator suffers more severely from increasing K, which is in line with the fact that the variance of the median y_{it} is larger than that of the mean y_{it} in our distributions of y_{it}.

A noteworthy constant in the discussion of results so far has been the equally good performance of DvS and BUC. This is remarkable as previous literature raised the concern that the DvS estimator could show difficulties when confronted with small samples for the different CML logit estimates. In the setup with K=10 and N=500 the last two CML logit estimators (k=9 and k=10) used on average about 137 and 78 individuals. DvS is only slightly (but statistically significantly) biased downwards. The last panel in Table 5.2 shows the results from a smaller sample of N=100 while maintaining K=10. This produces a difficult DGP for DvS, as only about 28 and 29 individuals are used in the CML logit estimations of k=9 and k=10. This resembles the situation in life satisfaction studies, where responses in lower categories are extremely infrequent (Ferrer-i-Carbonell and Frijters, 2004). Here we do find biases of -6% and -7% for DvS (margin of error at 99%: 1% and 2%, respectively). The BUC estimator in contrast remains as unbiased as in previous DGPs. FF, Median and Mean estimators also show little change and are as biased as with N=500.

The influence of the distribution of the data on the performance of the estimators is

addressed in the DGPs whose results are shown in Table 5.3. Again, the first panel repeats the results for the baseline case from Table 5.1. The next two panels —with headings 'bell-shaped y' and 'uniform y' — show results for different marginal distributions of y_{it}. I.e., all parameters from the baseline DGP are kept unchanged, except for thresholds κ which have been shifted to yield these distributions (cf. Fig. 5.1). These changes in y_{it} seem to have close to no impact on the performance of the estimators. Only CML logit estimators are affected in their precision. It is no surprise that, for given distribution of x, d, the more balanced the distribution of the dichotomized variable, the higher the precision of the resulting CML logit estimator. The last panel in Table 5.3 resets the thresholds to their baseline values and changes the distribution of the explanatory variables. The continuous x is now drawn from a log-normal distribution, standardized to have mean zero and unit variance; the binary d's new distribution is highly unbalanced with only 10% of observations having $d = 1$ on average. As before, the picture remains by and large the same: All estimators show higher standard deviations in this DGP, but the ranking is unchanged; CML logit estimators suffer the largest precision losses.

5.4 Application: Why are the unemployed so unhappy?

The preceding section documented the performance of different estimators for the fixed effects ordered logit model in simulations. In this section, the estimators are used to reestimate the effect of unemployment on life satisfaction using the dataset of Winkelmann and Winkelmann (1998). The data consists of a large sample from the German Socioeconomic Panel, totaling at 20,944 observations; the model includes 9 explanatory variables. With these values, the application provides a typical setting to which the estimators have been put to use and are likely to be applied in the future.

5.4.1 Data and specification

The sample consists of the first six consecutive waves of the German Socioeconomic Panel going from 1984 to 1989. It includes all observations of persons aged 20-64 years with

participation in at least two waves and non-missing responses for all variables of the model. These are 20,944 person-year observations corresponding to 4,261 individuals. Of these, 1,873 observations corresponding to 303 persons are discarded because they do not display any variation over time in their outcome variable, leaving the dataset with 19,079 observations corresponding to 3,958 individuals.

The outcome variable is satisfaction with life which is measured as the answer to the question *"How satisfied are you at present with your life as a whole?"*. The answer can be indicated in 11 ordered categories ranging from 0, "completely dissatisfied", to 10, "completely satisfied".

The key explanatory variables are a set of three dummy variables which indicate current labor market status: *Unemployed*, *Employed* and *Out of labor force*. These dummies exhaust the possible labor market status and are mutually exclusive, so *Employed* is used as the omitted reference category in the model.

Additional information about psychological costs of unemployment might be revealed through the length of the unemployment spell. Thus, the model contains the variables *Duration of unemployment* and *Squared duration of unemployment*.

Demographic control variables include marital status (*Married*), health status (*Good health*), age (*Age* and *Squared age*) and household income (in logarithms, *Log. household income*). We refer to the original source for comprehensive descriptions of data and specification.

5.4.2 Results

Estimation results are presented in Table 5.4. Every column depicts results of the same model for a different estimator. The first replicates the original results in Winkelmann and Winkelmann (1998, Table 4, column 2, p.11) who used a CML logit estimator dichotomized at the cutoff 8. This cutoff results in a distribution of the binary dependent variable which is about balanced with around 50% of the responses being equal or greater than 8. In

total 2,573 individuals cross this cutoff resulting in an estimating sample size of 12,980 observations.

To briefly summarize the results, the effect of unemployment is found to be both large and statistically significant; there is no effect of unemployment duration on life satisfaction, so there seems to be no mental adaptation process of unemployed persons to their status. Coefficients of socio-demographic variables display expected signs and magnitudes (cf. Clark and Oswald, 1994).

Moving to the right of the table, the next two columns correspond to results obtained using the DvS and BUC estimators, and the final three columns show results for FF, Mean and Median estimates. The most striking feature of the results as a whole is that the first three columns —which are based on consistent estimators— are remarkably similar, while they differ from the three last columns containing estimates from inconsistent estimators. The marginal effect of unemployment on latent life satisfaction is estimated to be around -1 when using CML logit, DvS or BUC; but it ranges only from -0.84 to -0.66 when using FF, Mean or Median estimators. Similarly, effects for marital status and age are estimated to be larger using either of the consistent estimators. Although estimation is not precise enough to reject equality of coefficients, these results clearly echo patterns from the Monte Carlo simulations. There is only one clear difference between consistent estimators. It relates to the coefficient of *Out of labor force*, which is -0.24 and insignificant for CML logit while being around -0.45 and significant for DvS and BUC. A potential explanation for this is that most of the changes in *Out of labor force* occur at levels of satisfaction lower than the cutoff used by the CML logit estimator, so that this information is lost to the CML logit estimation. DvS and BUC, on the other hand, use all 3,958 persons displaying some time variation in life satisfaction (for BUC the number of persons corresponds to the number of clusters; the number of individuals is the cross-sectional dimension of the "blown-up" or inflated sample).

All estimators display similarly sized standard errors. CML logit estimates are the

largest, but this difference is not pronounced. BUC's standard errors are slightly larger than DvS and the inconsistent estimators', the difference being minimal. There is another negligible difference between estimators. While Mean and Median estimators use the same number of persons as DvS and BUC for estimation (3,958), FF uses 9 individuals (or 0.2%) less. This means that for these 9 individuals, the smallest Hessians (remember that the smallest Hessian determines the dichotomizing cutoff for FF) were to be found for cutoffs which lead to no time variation in y.

5.5 Conclusions

This article studied extant estimators for the fixed effects ordered logit model and proposed a new one. All these estimators are based on CML binary logit estimation. Estimators most represented in the literature are characterized by selecting the dichotomizing cut-off point endogenously, i.e. as a function of the outcome of the dependent variable. In general, this will lead to inconsistency of the estimator, a result which was extensively documented in Monte Carlo simulations. The consistent estimators, Das and van Soest's (1999) minimum distance estimator and our BUC estimator, clearly outperformed simple CML logit estimation in terms of efficiency. Their performance in several DGPs of the Monte Carlo simulations and in a large-scale application using survey data from Germany indicates that they are recommendable for applied work. If the ordered dependent variable displays extremely low responses for some categories, our simulation evidence suggest that BUC estimation is preferable.

References

Andersen, Erling B. (1970), "Asymptotic Properties of Conditional Maximum-likelihood Estimators", *Journal of the Royal Statistical Society, Series B (Methodological)*, **32**, 283-301.

Booth, Alison L., and Jan C. van Ours (2008), "Job Satisfaction and Family Happiness: The Part-time Work Puzzle", *Economic Journal*, **118**, F77-F99.

Chamberlain, Gary (1980), "Analysis of covariance with qualitative data", *Review of Economic Studies*, **47**, 225-238.

Clark, Andrew E. and Andrew J. Oswald (1994), "Unhappiness and unemployment", *Economic Journal*, **104**, 648-659.

Das, Marcel, and Arthur van Soest (1999), "A panel data model for subjective information on household income growth", *Journal of Economic Behavior & Organization*, **40**, 409-426.

D'Addio, Anna Cristina, Tor Eriksson and Paul Frijters (2007), "An Analysis of the Determinants of Job Satisfaction when Individuals' Baseline Satisfaction Levels May Differ", *Applied Economics*, **39**, 2413-2423.

Ferrer-i-Carbonell, Ada, and Paul Frijters (2004), "How important is methodology for the estimates of the determinants of happiness?", *Economic Journal*, **114**, 641-659.

Frijters, Paul, John P. Haisken-DeNew and Michael A. Shields (2004 a), "Investigating the Patterns and Determinants of Life Satisfaction in Germany Following Reunification", *Journal of Human Resources*, **39**, 649-674.

Frijters, Paul, John P. Haisken-DeNew and Michael A. Shields (2004 b), "Money Does Matter! Evidence from Increasing Real Income and Life Satisfaction in East Germany Following Reunification", *American Economic Review*, **94**, 730-740.

Frijters, Paul, John P. Haisken-DeNew and Michael A. Shields (2005), "The causal effect of income on health: Evidence from German reunification", *Journal of Health Economics*, **24**, 997-1017.

Greene, William H. (2004), "The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects", *Econometrics Journal*, **7**, 98-119.

Jones, Andrew M., and Stephanie Schurer, "How does heterogeneity shape the socioeconomic gradient in health satisfaction?", forthcoming in *Journal of Applied Econometrics*, published online December 14 2009, DOI: 10.1002/jae.1133

Kassenboehmer, Sonja C., and John P. Haisken-DeNew (2009), "You're Fired! The Causal Negative Effect of Unemployment on Life Satisfaction", *Economic Journal*, **119**, 448-462.

White, Halbert (1982), "Maximum likelihood estimation of misspecified models", *Econometrica*, **50**, 1-25.

Winkelmann, Liliana, and Rainer Winkelmann (1998), "Why are the unemployed so unhappy? Evidence from panel data", *Economica*, **65**, 1-15.

A Implementing the BUC estimator in Stata

To perform BUC estimation in Stata, run the following code, replacing `ivar yvar xvar` in the last program line as follows:

`ivar` is the individual identifier,
`yvar` is the ordered dependent variable, and
`xvars` is the list of explanatory variables.

```
capture program drop feologit_buc
program feologit_buc, eclass
    version 10
    gettoken gid 0: 0
    gettoken y x: 0
    tempvar iid id cid gidcid dk

    qui sum ‘y’
    local lk= r(min)
    local hk= r(max)
    bys ‘gid’: gen ‘iid’=_n
    gen long ‘id’=‘gid’*100+‘iid’
    expand ‘=‘hk’-‘lk’’
    bys ‘id’: gen ‘cid’=_n
    qui gen long ‘gidcid’= ‘gid’*100+‘cid’
    qui gen ‘dk’= ‘y’>=‘cid’+1

    clogit ‘dk’ ‘x’, group(‘gidcid’) cluster(‘gid’)
end

feologit_buc ivar yvar xvars
```

B Inconsistency of estimators with endogenous cut-offs for T=3, K=3

Here we analytically examine consistency of the estimators in a particular setup: T=3, K=3, $x_i = x$ for all i. We show inconsistency of the mean estimator in this case. Thus, the mean estimator is inconsistent, in general. This setup is particularly convenient for two reasons. First, it is simple and tractable. Second, in this setup the mean estimator is equal to the median estimator, thus extending the inconsistency result to the median estimator. Finally, for particular values of x and β, the mean estimator is also equivalent to Ferrer-i-Carbonell and Frijters' (2004) estimator (FF), showing that the FF estimator, too, is inconsistent, in general.

The x's change within an individual over time, but only the individual fixed effect α_i is allowed to change between individuals. We treat x_i and α_i as fixed. If a particular estimator is consistent for arbitrary fixed x's and α's, it is also consistent for varying x's and α's.

B.1 Probability limit of the score

First we derive the probability limit of the score of the estimators to be examined. These are the CML logit estimators dichotomized at 2 and at 3, the mean, median and FF estimators. Since all estimators have the same score structure and differ only by the dichotomization rule, we index estimators by $c \in \{k = 2, k = 3, \mathrm{Mn}, \mathrm{Md}, \mathrm{FF}\}$, respectively. Then, the probability limit of estimator's c score is

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} s_i^c(b) &= x' \operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\left[d_i^c - \sum_{\mathcal{I}(j)=\mathcal{I}(d_i^c)} j \frac{\exp(j'xb)}{\sum_{\mathcal{I}(l)=\mathcal{I}(d_i)}\exp(l'xb)}\right] \\
&= x' \operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{a=1}^{2} \mathbb{1}(\mathcal{I}(d_i^c)=a)\left(d_i^c - \sum_{\mathcal{I}(j)=a} j \frac{\exp(j'xb)}{\sum_{\mathcal{I}(l)=a}\exp(l'xb)}\right)\right] \\
&= x' \sum_{a=1}^{2} \Pr(\mathcal{I}(d^c)=a)\left[\mathrm{E}(d^c|\,\mathcal{I}(d^c)=a) - \sum_{\mathcal{I}(j)=a} j \frac{\exp(j'xb)}{\sum_{\mathcal{I}(l)=a}\exp(l'xb)}\right],
\end{aligned}$$

where d_i^c is the binary dependent variable obtained by using dichotomizing rule c. $\mathbb{1}(z)$ denotes the indicator function (equal to 1 if z is true, 0 otherwise), and $\mathcal{I}(j) = \sum_t \mathbb{1}(j_t = 1)$. I.e., $\mathcal{I}(j)$ is the function that returns the number of elements in j that are equal to one. We use $\sum_{\mathcal{I}(j)=a} f(j)$ to denote the sum of $f(j)$ over all vectors j satisfying $\mathcal{I}(j) = a$.

Setting the score to zero yields an implicit function for estimator c ($\hat{\beta}^c$). If for all relevant values of a (here: 1,2; a=0 and a=3 do not contribute to the score) it holds that

$$\mathrm{E}\left(d^c | \mathcal{I}(d^c) = a\right) - \sum_{\mathcal{I}(j)=a} j \frac{\exp(j' x\beta)}{\sum_{\mathcal{I}(l)=a} \exp(l' x\beta)} = 0 \tag{B.1}$$

it follows that estimator c is consistent. If this is the case, the score is zero if and only if $b = \beta$ because the score is monotonic in b. I.e., if we show that the conditional expectation of the dependent variable dichotomized using rule c is

$$\mathrm{E}\left(d^c | \mathcal{I}(d^c) = a\right) = \sum_{\mathcal{I}(j)=a} j \frac{\exp(j' x\beta)}{\sum_{\mathcal{I}(l)=c} \exp(l' x\beta)} \quad \forall a \in 1, 2 \tag{B.2}$$

then estimator c is consistent.

To derive $\mathrm{E}\left(d^c | \mathcal{I}(d^c)\right)$ for the estimators in question, it is helpful to be aware of some simple ordered logit formulas

$$\frac{\Pr(y_{it} \geq k)}{\Pr(y_{it} < k)} = \frac{1 - \frac{\exp(\kappa_{k-1} - x_t'\beta - \alpha_1)}{1 + \exp(\kappa_{k-1} - x_t'\beta - \alpha_1)}}{\frac{\exp(\kappa_{k-1} - x_t'\beta - \alpha_1)}{1 + \exp(\kappa_{k-1} - x_t'\beta - \alpha_1)}} = \frac{\exp(x_t'\beta + \alpha_1)}{\exp(\kappa_{k-1})} = \frac{\exp(x_t'\beta)}{\exp(\kappa_{k-1}) / \exp(\alpha_i)} \tag{B.3}$$

$$\frac{\Pr(y_{it} = 3)}{\Pr(y_{it} = 2)} = \frac{1 - \frac{\exp(\kappa_2 - x_t'\beta - \alpha_i)}{1 + \exp(\kappa_2 - x_t'\beta - \alpha_i)}}{\frac{\exp(\kappa_2 - x_t'\beta - \alpha_i)}{1 + \exp(\kappa_2 - x_t'\beta - \alpha_i)} - \frac{\exp(\kappa_1 - x_t'\beta - \alpha_i)}{1 + \exp(\kappa_1 - x_t'\beta - \alpha_i)}} = \frac{\exp(x_t'\beta) + \exp(\kappa_1) / \exp(\alpha_i)}{\exp(\kappa_2) / \exp(\alpha_i) - \exp(\kappa_1) / \exp(\alpha_i)} \tag{B.4}$$

For notational ease we use, for example, $\Pr(1, > 1, \geq 2)$ to denote $\Pr(y_1 = 1, y_2 > 1, y_3 \geq 2)$. Note that the y_t's within an individual are independent if we either condition on α_i and x_i, or threat α_i and x_i as fixed: $\Pr(y_1 = 1, y_2 > 1, y_3 \geq 2) = \Pr(y_1 = 1) \cdot \Pr(y_2 > 1) \cdot \Pr(y_3 \geq 2)$.

B.2 Consistency of estimators with exogenous cutoff

We begin by showing that estimators dichotomizing at a fixed cutoff point (k=2,3) are consistent in this setup. The procedure is as follow: We derive $\mathrm{E}(d^k | \mathcal{I}(d^k) = a)$, for $a = 1, 2$. If both expressions are equal to the right hand side of (B.2) for each a, we have shown that the estimator is consistent.

$a = 1$

$$
\begin{aligned}
&\mathrm{E}(d^k | \mathcal{I}(d^k) = 1) \\
&= \frac{\binom{1}{0}{0} \Pr(\geq k, < k, < k) + \binom{0}{1}{0} \Pr(< k, \geq k, < k) + \binom{0}{0}{1} \Pr(< k, < k \geq k)}{\Pr(\geq k, < k, < k) + \Pr(< k, \geq k, < k) + \Pr(< k, < k \geq k)} \\
&= \frac{\binom{1}{0}{0} \frac{\Pr(y_1 \geq k)}{\Pr(y_1 < k)} + \binom{0}{1}{0} \frac{\Pr(y_2 \geq k)}{\Pr(y_2 < k)} + \binom{0}{0}{1} \frac{\Pr(y_3 \geq k)}{\Pr(y_3 < k)}}{\sum_{t=1}^{3} \frac{\Pr(y_t \geq k)}{\Pr(y_t < k)}} \\
&= \binom{1}{0}{0} \frac{\exp(x_1'\beta)}{\sum_{t=1}^{3} \exp(x_t'\beta)} + \binom{0}{1}{0} \frac{\exp(x_2'\beta)}{\sum_{t=1}^{3} \exp(x_t'\beta)} + \binom{0}{0}{1} \frac{\exp(x_3'\beta)}{\sum_{t=1}^{3} \exp(x_t'\beta)}
\end{aligned} \tag{B.5}
$$

where $k \in \{2, 3\}$ denotes the fixed cutoff. The last expression is equal to the right hand side of (B.2) for $a = 1$.

$a = 2$

$$
\begin{aligned}
&\mathrm{E}(d^k | \mathcal{I}(d^k) = 2) \\
&= \frac{\binom{0}{1}{1} \Pr(< k, \geq k, \geq k) + \binom{1}{0}{1} \Pr(\geq k, < k, \geq k) \binom{1}{1}{0} \Pr(\geq k, \geq k, < k)}{\Pr(< k, \geq k, \geq k) + \Pr(\geq k, < k, \geq k) + \Pr(\geq k, \geq k, < k)} \\
&= \frac{\binom{0}{1}{1} \frac{\Pr(y_1 < k)}{\Pr(y_1 \geq k)} + \binom{1}{0}{1} \frac{\Pr(y_2 < k)}{\Pr(y_2 \geq k)} + \binom{1}{1}{0} \frac{\Pr(y_3 < k)}{\Pr(y_3 \geq k)}}{\sum_{t=1}^{3} \frac{\Pr(y_t < k)}{\Pr(y_t \geq k)}} \\
&= \frac{\binom{0}{1}{1} \exp(x_1'\beta)^{-1} + \binom{1}{0}{1} \exp(x_2'\beta)^{-1} + \binom{1}{1}{0} \exp(x_3'\beta)^{-1}}{\sum_{t=1}^{3} \exp(x_t'\beta)^{-1}} \\
&= \binom{0}{1}{1} \frac{\exp(x_2'\beta + x_3'\beta)}{\sum_t \exp(\sum_{m \neq t} x_m'\beta)} + \binom{1}{0}{1} \frac{\exp(x_1'\beta + x_3'\beta)}{\sum_t \exp(\sum_{m \neq t} x_m'\beta)} + \binom{1}{1}{0} \frac{\exp(x_1'\beta + x_2'\beta)}{\sum_t \exp(\sum_{m \neq t} x_m'\beta)}
\end{aligned} \tag{B.6}
$$

The last expression is equal to the right hand side of (B.2) for $a = 2$. Because a can be only either 1 or 2, we have shown that the conditional logit estimator with a fixed cutoff is consistent in this setup.

B.3 Inconsistency of estimators with endogenous cutoff

Now we show that estimators with endogenous cutoff are inconsistent, in general. It is sufficient to show this for the mean estimator, because with K=3 and T=3, mean and median estimators produce the same dichotomized binary variable. Furthermore, for some

values of x and β, the mean estimator will produce the same dichotomized binary variable than the FF estimator. We give examples of such cases at the end of this section.

To study the mean estimator, we further partition the score into mutually exclusive sets.

$$\begin{aligned}\mathrm{E}(d^{\mathrm{Mn}}|\,\mathcal{I}(d^{\mathrm{Mn}})=a) = \Pr(\mathcal{I}(y)=v|\,\mathcal{I}(d^{\mathrm{Mn}})=a)\cdot \mathrm{E}(d^{\mathrm{Mn}}|\,\mathcal{I}(d^{\mathrm{Mn}})=a,\,\mathcal{I}(y)=v)\\ + \Pr(\mathcal{I}(y)\neq v|\,\mathcal{I}(d^{\mathrm{Mn}})=a)\cdot \mathrm{E}(d^{\mathrm{Mn}}|\,\mathcal{I}(d^{\mathrm{Mn}})=a,\,\mathcal{I}(y)\neq v)\end{aligned} \tag{B.7}$$

The first set consists of cases with v 1's in the y-vector. The second set consists of the remaining cases.

The procedure is the following: First we consider $\mathrm{E}(d^{\mathrm{Mn}}|\,\mathcal{I}(d^{\mathrm{Mn}})=1)$. We will partition the expectation in those cases with $\mathcal{I}(y)=2$ —for instance, y=(1,2,1)' or y=(3,1,1)'— and those with $\mathcal{I}(y)\neq 2$. We show that the expectation of the first set has the desired form (B.2), while the second set does not. Therefore, the score contibution evaluated at the true β is not zero for $a=1$ if we dichotomize at the individual mean. Then we repeat the analysis for $a=2$ and $\mathcal{I}(y)=1$, finding the same pattern. Finally, we will show that, in general, the two score contributions which are different from (B.2) do not add to zero; this implies that the mean estimator is not consistent.

$a=1$

Consider the case when the vector d^{Mn} has one 1 and two 0's ($a=1$) and the associated y-vector has two 1's ($\mathcal{I}(y)=2$).

$$\begin{aligned}&\mathrm{E}(d^{\mathrm{Mn}}|\,\mathcal{I}(d^{\mathrm{Mn}})=1,\,\mathcal{I}(y)=2)\\ &= \frac{\binom{1}{0}_{0}\Pr(\geq 2,1,1)+\binom{0}{1}_{0}\Pr(1,\geq 2,1)+\binom{0}{0}_{1}\Pr(1,1,\geq 2)}{\Pr(\geq 2,1,1)+\Pr(1,\geq 2,1)+\Pr(1,1,\geq 2)}\\ &= \frac{\binom{1}{0}_{0}\Pr(\geq 2,<2,<2)+\binom{0}{1}_{0}\Pr(<2,\geq 2,<2)+\binom{0}{0}_{1}\Pr(<2,<2,\geq 2)}{\Pr(\geq 2,<2,<2)+\Pr(<2,\geq 2,<2)+\Pr(<2,<2,\geq 2)}\\ &= \frac{\binom{1}{0}_{0}\frac{\Pr(y_1\geq 2)}{\Pr(y_1<2)}+\binom{0}{1}_{0}\frac{\Pr(y_2\geq 2)}{\Pr(y_2<2)}+\binom{0}{0}_{1}\frac{\Pr(y_3\geq 2)}{\Pr(y_3<2)}}{\sum_{t=1}^{3}\frac{\Pr(y_t\geq 2)}{\Pr(y_t<2)}}\end{aligned} \tag{B.8}$$

The last expression is equal to right hand side of (B.2). Now we look at the remaining part of $\mathrm{E}(d^{\mathrm{Mn}}|\,\mathcal{I}(d^{\mathrm{Mn}})=1)$. The only y-vectors satisfying $\mathcal{I}(y)\neq 2$ and $\mathcal{I}(d^{\mathrm{Mn}})=1$ are cases

with two 2's and one 3.

$$
\begin{aligned}
&\mathrm{E}(d^{\mathrm{Mn}}|\mathcal{I}(d^{\mathrm{Mn}})=1,\mathcal{I}(y)\neq 2)\\
&=\frac{\binom{1}{0}{0}\Pr(3,2,2)+\binom{0}{1}{0}\Pr(2,3,2)+\binom{0}{0}{1}\Pr(2,2,3)}{\Pr(3,2,2)+\Pr(2,3,2)+\Pr(2,2,3)}\\
&=\frac{\binom{1}{0}{0}\frac{\Pr(y_1=3)}{\Pr(y_1=2)}+\binom{0}{1}{0}\frac{\Pr(y_2=3)}{\Pr(y_2=2)}+\binom{0}{0}{1}\frac{\Pr(y_3=3)}{\Pr(y_3=2)}}{\sum_{t=1}^{3}\frac{\Pr(y_t=3)}{\Pr(y_t=2)}}\\
&=\frac{\binom{1}{0}{0}(\exp(x_1'\beta)+\varkappa_1))+\binom{0}{1}{0}(\exp(x_2'\beta)+\varkappa_1)+\binom{0}{0}{1}(\exp(x_3'\beta)+\varkappa_1)}{\sum_{t=1}^{3}(\exp(x_t'\beta)+\varkappa_1)},
\end{aligned}
\tag{B.9}
$$

where $\varkappa_1 \equiv \exp(\kappa_1)\mathrm{E}(\exp(-\alpha_i))$. This expression is only equal to the right hand side of (B.2) if $\exp(\kappa_1)=0$. This is only possible if κ_1 goes to minus infinity which means that the probability if $y_{it}=1$ is zero (i.e., this is the limiting case with two categories: K=2). Thus, the score contribution for $a=1$ evaluated at $b=\beta$ is not equal to zero if we dichotomize at the individual mean.

$a=2$

Now we consider cases where the number of 1's in the d^{Mn}-vector is 2 ($a=2$). We divide these cases into those satisfying $\mathcal{I}(y)=1$ and the rest. If we dichotomize at the individual mean, the only y-vectors for which $\mathcal{I}(y)=1$ and $\mathcal{I}(d)=2$ are those with one 2 and two 3's.

$$
\begin{aligned}
&\mathrm{E}(d^{\mathrm{Mn}}|\mathcal{I}(d^{\mathrm{Mn}})=2,\mathcal{I}(y)=1)\\
&=\frac{\binom{0}{1}{1}\Pr(1,\geq 2,\geq 2)+\binom{1}{0}{1}\Pr(\geq 2,1,\geq 2)+\binom{1}{1}{0}\Pr(\geq 2,\geq 2,1)}{\Pr(1,\geq 2,\geq 2)+\Pr(\geq 2,1,\geq 2)+\Pr(\geq 2,\geq 2,1)}\\
&=\frac{\binom{0}{1}{1}\frac{\Pr(y_1<2)}{\Pr(y_1\geq 2)}+\binom{1}{0}{1}\frac{\Pr(y_2<2)}{\Pr(y_2\geq 2)}+\binom{1}{1}{0}\frac{\Pr(y_3<2)}{\Pr(y_3\geq 2)}}{\frac{\Pr(y_1<2)}{\Pr(y_1\geq 2)}+\frac{\Pr(y_2<2)}{\Pr(y_2\geq 2)}+\frac{\Pr(y_3<2)}{\Pr(y_3\geq 2)}}
\end{aligned}
\tag{B.10}
$$

This is equivalent to the right hand side of (B.2).

$$
\begin{aligned}
&\mathrm{E}(d^{\mathrm{Mn}}|\,\mathcal{I}(d^{\mathrm{Mn}})=2,\,\mathcal{I}(y)\neq 1)\\
&\quad=\frac{\binom{0}{1}{\scriptstyle 1}\Pr(2,3,3)+\binom{1}{0}{\scriptstyle 1}\Pr(3,2,3)+\binom{1}{1}{\scriptstyle 0}\Pr(3,3,2)}{\Pr(2,3,3)+\Pr(3,2,3)+\Pr(3,3,2)}\\
&\quad=\frac{\binom{0}{1}{\scriptstyle 1}\frac{\Pr(y_1=2)}{\Pr(y_1=3)}+\binom{1}{0}{\scriptstyle 1}\frac{\Pr(y_2=2)}{\Pr(y_2=3}+\binom{1}{1}{\scriptstyle 0}\frac{\Pr(y_3=2)}{\Pr(y_3=3)}}{\sum_{t=1}^{3}\frac{\Pr(y_t=2)}{\Pr(y_t=3)}}\\
&\quad=\frac{\binom{0}{1}{\scriptstyle 1}\left(\exp(x_1'\beta)+\varkappa_1\right)^{-1}+\binom{1}{0}{\scriptstyle 1}\left(\exp(x_2'\beta)+\varkappa_1\right)^{-1}+\binom{1}{1}{\scriptstyle 0}\left(\exp(x_3'\beta)+\varkappa_1\right)^{-1}}{\sum_{t=1}^{3}\left(\exp(x_t'\beta)+\varkappa_1\right)^{-1}}
\end{aligned}
\tag{B.11}
$$

This expression is only equivalent to the right hand side of (B.2) if $\exp(\kappa_1)$ vanishes. Thus the score contibution for $a = 2$ evaluated at $\hat{\beta} = \beta$ is not equal to zero if we dichotomize at the individual mean.

If, for instance, $\varkappa_1$=1, $\beta = 1$, and x_t are scalar with $x_t = \ln(t)$, it is easy to verify that both score contributions (B.1) for $a = 1$ and $a = 2$ are negative. Thus, in general, the two non-zero score contributions do not cancel out, because the probability weights are necessarily positive. This implies that the mean estimator is inconsistent. Moreover, it is easy to verify that mean and FF estimator coincide in this DGP. This implies that the FF estimator is inconsistent, in general, too.

Table 5.1: Monte Carlo simulation results (1,000 replications): Baseline scenario

	$\hat{\beta}_x$			$\hat{\beta}_d$		
Estimators	M	SD	SE	M	SD	SE
DvS	1.00	0.07	0.07	0.99	0.11	0.11
BUC	1.00	0.07	0.07	1.00	0.12	0.12
FF	0.93	0.07	0.07	0.92	0.12	0.12
Median	0.94	0.07	0.07	0.94	0.12	0.12
Mean	0.96	0.07	0.07	0.95	0.12	0.12
$y \geq 2$	1.00	0.09	0.09	1.00	0.15	0.15
$y \geq 3$	1.01	0.09	0.09	1.00	0.15	0.16
$y \geq 4$	1.01	0.12	0.11	1.00	0.20	0.20
$y \geq 5$	1.03	0.18	0.18	1.02	0.32	0.32

Notes: $\beta_x = \beta_d = 1$. Columns labeled M contain the mean of the estimated coefficients over all replications, columns SD the standard deviation of the estimated coefficients, and columns SE contain the mean of the estimated standard errors. Baseline scenario is N=500, T=4, K=5, $x \sim Normal(0, 1)$, $d \sim Bernoulli(0.5)$, skewed distribution for y.

Table 5.2: Monte Carlo simulation results (1,000 replications): The effects of increasing N, T and K

	Baseline scenario: N=500, T=4, K=5				*N=1,000*				*T=8*				*K=10*				*K=10, N=100*			
	$\hat{\beta}_x$		$\hat{\beta}_d$		$\hat{\beta}_x$		$\hat{\beta}_d$		$\hat{\beta}_x$		$\hat{\beta}_d$		$\hat{\beta}_x$		$\hat{\beta}_d$		$\hat{\beta}_x$		$\hat{\beta}_d$	
Estimators	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
DvS	1.00	0.07	0.99	0.11	1.00	0.05	1.00	0.08	1.00	0.04	1.00	0.08	0.99	0.06	0.98	0.10	0.93	0.14	0.94	0.24
BUC	1.00	0.07	1.00	0.12	1.00	0.05	1.00	0.08	1.00	0.04	1.00	0.08	1.00	0.06	1.00	0.11	1.00	0.14	1.00	0.24
FF	0.93	0.07	0.92	0.12	0.93	0.05	0.93	0.09	0.96	0.04	0.96	0.08	0.86	0.06	0.86	0.11	0.86	0.13	0.86	0.26
Median	0.94	0.07	0.94	0.12	0.94	0.05	0.94	0.09	0.97	0.04	0.97	0.08	0.88	0.06	0.87	0.11	0.88	0.14	0.88	0.25
Mean	0.96	0.07	0.95	0.12	0.96	0.05	0.96	0.09	0.98	0.05	0.97	0.08	0.94	0.07	0.93	0.11	0.95	0.15	0.94	0.26
$y \geq 2$	1.00	0.09	1.00	0.15	1.00	0.06	1.00	0.11	1.00	0.05	1.00	0.10	1.01	0.10	1.00	0.18	1.04	0.23	1.02	0.41
$y \geq 3$	1.01	0.09	1.00	0.15	1.00	0.06	1.00	0.11	1.00	0.06	1.00	0.10	1.01	0.09	1.00	0.15	1.02	0.20	1.01	0.34
$y \geq 4$	1.01	0.12	1.00	0.20	1.00	0.08	1.00	0.14	1.00	0.07	1.00	0.13	1.00	0.08	1.00	0.14	1.02	0.19	1.01	0.33
$y \geq 5$	1.03	0.18	1.02	0.32	1.01	0.13	1.01	0.22	1.01	0.11	1.01	0.20	1.01	0.09	1.00	0.15	1.01	0.20	1.00	0.34

Notes: $\beta_x = \beta_d = 1$. Columns labeled M contain the mean of the estimated coefficients over all replications, columns SD the standard deviation of the estimated coefficients. Baseline scenario is N=500, T=4, K=5, $x \sim Normal(0,1)$, $d \sim Bernoulli(0.5)$, skewed distribution for y. Departures from baseline scenario are noted in the top row.

Table 5.3: Monte Carlo simulation results (1,000 replications): Changing the distributions of y, x and d

	Baseline: skewed y, $x \sim N(0,1)$, $d \sim B(0.5)$				*bell-shaped* y				*uniform* y				$x \sim LN(0,1), d \sim B(0.1)$			
	$\hat{\beta}_x$		$\hat{\beta}_d$		$\hat{\beta}_x$		$\hat{\beta}_d$		$\hat{\beta}_x$		$\hat{\beta}_d$		$\hat{\beta}_x$		$\hat{\beta}_d$	
Estimators	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
DvS	1.00	0.07	0.99	0.11	1.00	0.06	0.99	0.11	0.99	0.06	0.99	0.11	0.99	0.08	0.99	0.17
BUC	1.00	0.07	1.00	0.12	1.00	0.06	1.00	0.11	1.00	0.06	1.00	0.11	1.00	0.08	1.00	0.17
FF	0.93	0.07	0.92	0.12	0.93	0.07	0.93	0.12	0.91	0.07	0.90	0.12	0.91	0.08	0.91	0.18
Median	0.94	0.07	0.94	0.12	0.94	0.07	0.93	0.12	0.92	0.07	0.91	0.11	0.94	0.09	0.94	0.18
Mean	0.96	0.07	0.95	0.12	0.95	0.07	0.95	0.12	0.94	0.07	0.94	0.11	0.96	0.09	0.95	0.18
$y \geq 2$	1.00	0.09	1.00	0.15	1.01	0.14	1.02	0.23	1.01	0.10	1.00	0.18	1.01	0.13	1.00	0.23
$y \geq 3$	1.01	0.09	1.00	0.15	1.01	0.09	1.00	0.16	1.00	0.09	1.00	0.15	1.01	0.11	1.01	0.22
$y \geq 4$	1.01	0.12	1.00	0.20	1.01	0.09	1.00	0.15	1.01	0.09	1.00	0.15	1.01	0.12	1.01	0.30
$y \geq 5$	1.03	0.18	1.02	0.32	1.01	0.13	1.01	0.23	1.01	0.10	1.00	0.18	1.03	0.19	1.01	0.57

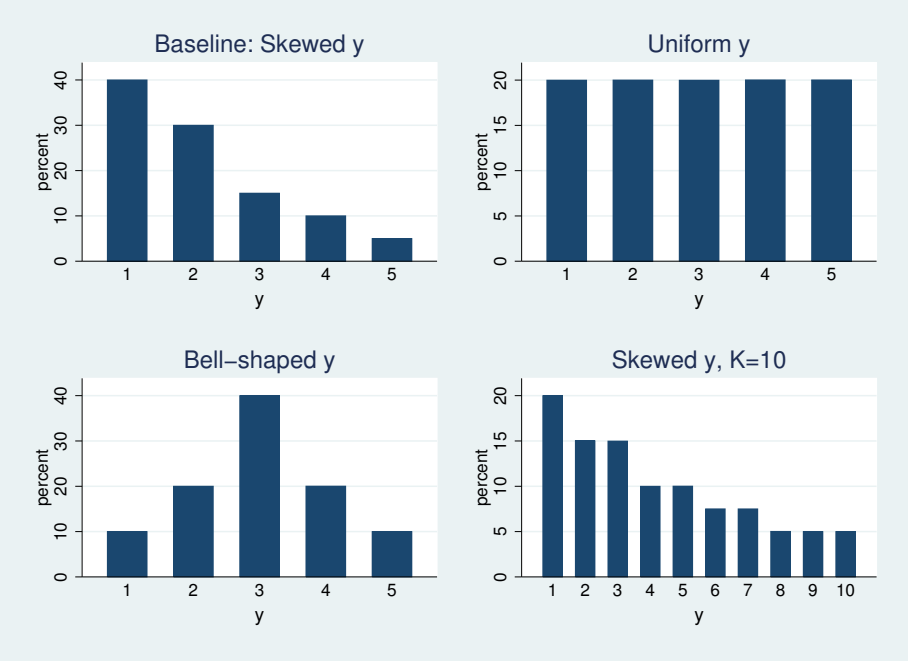
Notes: $\beta_x = \beta_d = 1$. Columns labeled M contain the mean of the estimated coefficients over all replications, columns SD the standard deviation of the estimated coefficients. Baseline scenario is N=500, T=4, K=5, $x \sim Normal(0,1)$, $d \sim Bernoulli(0.5)$, skewed distribution for y (first four columns). Departures from baseline scenario are noted in the top row. $LN()$ stands for log-normal distribution.

Table 5.4: Fixed effects ordered logit estimates of life satisfaction

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.: *Life Satisfaction*	$y \geq 8$	DvS	BUC	FF	Mean	Median
Unemployed	-0.96**	-0.98**	-1.03**	-0.77**	-0.84**	-0.66**
	(0.20)	(0.14)	(0.16)	(0.15)	(0.15)	(0.15)
Out of labor force	-0.24	-0.42**	-0.45**	-0.25**	-0.25**	-0.25**
	(0.12)	(0.09)	(0.11)	(0.09)	(0.10)	(0.09)
Duration of unemployment	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Squared duration of unemp. $\times 10,000^{-1}$	0.60	2.44	2.75	3.18	2.17	2.12
	(2.79)	(1.56)	(2.30)	(1.87)	(1.88)	(1.86)
Married	0.67**	0.52**	0.56**	0.37**	0.39**	0.37**
	(0.12)	(0.09)	(0.11)	(0.09)	(0.09)	(0.09)
Good health	0.34**	0.33**	0.36**	0.24**	0.29**	0.24**
	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Age	-0.12**	-0.12**	-0.12**	-0.12**	-0.11**	-0.12**
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Squared age $\times 100^{-1}$	-0.84	-2.46	-1.15	-1.30	-2.91	-1.58
	(4.27)	(3.24)	(3.82)	(3.36)	(3.38)	(3.35)
Log. household income	0.13*	0.12**	0.13*	0.10*	0.10*	0.10*
	(0.06)	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)
$\log L$	-4,996	—	-21,802	-8,003	-7,911	-8,054
Observations	12,980	—	59,535	19,053	19,071	19,071
Individuals	2,573	3,958	11,864	3,949	3,958	3,958
Clusters	—	—	3,958	—	—	—

Notes: Data Source GSOEP, waves 1984-1989. * statistical significance at 5% level, ** statistical significance at 1% level. Observations denotes the number of person-year observations in estimation sample; Individuals denotes number of unique persons in estimation sample; Clusters denotes the number of groups in cluster-robust standard errors.

Figure 5.1: Marginal distribution of y in Monte Carlo experiments



Curriculum Vitae

Personal Information

Date of Birth: November 16, 1981

Citizenship: Swiss, Bolivian

Education

Doctoral studies in Economics, University of Zurich, Switzerland, April 2007 – April 2011

Master of Arts UZH, University of Zurich, Switzerland, April 2007

Research and teaching fields

Primary: Microeconometrics, applied econometrics.

Secondary: International trade, labor economics, health economics.

Publications

Egger P., M. Larch, K.E. Staub and R. Winkelmann (2011), "The Trade Effects of Endogenous Preferential Trade Agreements," forthcoming in *American Economic Journal: Economic Policy*.

Boes S., K.E. Staub and R. Winkelmann (2010), "Relative Status and Satisfaction," *Economics Letters*, **109**(3), 168-170.

Staub K.E. (2009), "Simple Tests for Exogeneity of a Binary Explanatory Variable in Count Data Regression Models," *Communications in Statistics: Simulation and Computation*, **38**(9), 1834-1855.

Research Papers

Baetschmann G., K.E. Staub and R. Winkelmann (2011), "Consistent estimation of the fixed effects ordered logit model," ECON Working Paper No. 4, Department of Economics, University of Zurich.

Staub K.E. (2010), "A causal interpretation of extensive and intensive margin effects in generalized Tobit models," SOI Working Paper No. 1012, Department of Economics, University of Zurich.

Staub K.E. and R. Winkelmann (2009), "Quasi-Likelihood Estimation of Zero-Inflated Count Data Models," revised version: March 2010, SOI Working Paper No. 0908, Department of Economics, University of Zurich.